MAIA

ERASMUS MUNDUS

JOINT MASTER IN MEDICAL IMAGING AND APPLICATIONS

Joint Master in Medical Imaging and Applications

Master Thesis Proceedings

Promotion 2017-19

www.maiamaster.org

MAIA

An international programme by the University of Girona (Spain), the University of Bourgogne (France) and the University of Cassino (Italy) funded by Erasmus + Programme.

Erasmus+ Universitat de Girona uB Université de Bourgogne

Published by the MAIA Master

www.maiamaster.org

This document is a compendium of the master thesis works developed by the students of the Joint Master Degree in Medical Imaging and Applications. Therefore, each work is independent on the other, and you should cite it individually as the final master degree report of the first author of each paper (Student name; title of the report; MAIA MSc Thesis; 2019).

Editorial

Computer aided applications for early detection and diagnosis, histopathological image analysis, treatment planning and monitoring, as well as robotised and guided surgery will positively impact health care during the new few years. The scientific community needs of prepared entrepreneurships with a proper ground to tackle these topics. The Joint Master Degree in Medical Imaging and Applications (MAIA) was born with the aim to fill this gap, offering highly skilled professionals with a depth knowledge on computer science, artificial intelligence, computer vision, medical robotics, and transversal topics.

The MAIA master is a two-years joint master degree (120 ECTS) between the Université de Bourgogne (uB, France), the Università degli studi di Cassino e del Lazio Meridionale (UNICLAM, Italy), and the Universitat de Girona (UdG, Spain), being the latter the coordinating institution. The program is supported by associate partners, that help in the sustainability of the program, not necessarily in economical terms, but in contributing in the design of the master, offering master thesis or internships, and expanding the visibility of the master. Moreover, the program is recognised by the European Commission for its academic excellence and is included in the list of Erasmus Mundus Joint Master Degrees under the Erasmus+ programme.

This document shows the outcome of the master tesis research developed by the MAIA students during the last semester, where they put their learnt knowledge in practice for solving different problems related with medical imaging. This include fully automatic anatomical structures segmentation, abnormality detection algorithms in different imaging modalities, biomechanical modelling, development of applications to be clinically usable, or practical components for integration into clinical workflows. We sincerely think that this document aims at further enhancing the dissemination of information about the quality of the master and may be of interest to the scientific community and foster networking opportunities amongst MAIA partners.

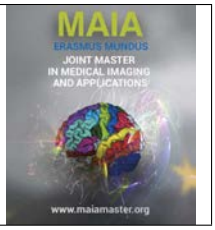
We finally want to thank and congratulate all the students for their effort done during this last semester of the Joint Master Degree in Medical Imaging and Applications.

MAIA Master Academic and Administrative Board

Contents

Medical Imaging and Applications

Master Thesis, June 2019



GAN (Generative Adversarial Networks) for realistic data augmentation and lesion simulation in x-ray breast imaging

Basel Alyafi, **Supervisors:** Robert Marti, Oliver Diaz

Universitat de Girona, 17003, Girona, Spain

Abstract

Early detection of breast cancer has a major contribution to curability, and this importance increases when using non-invasive solutions as mammographic images. Supervised deep learning methods have played a great role in object detection in computer vision, but it suffers from a limiting property; the need to huge labelled data. This becomes stricter when it comes to medical datasets which have high-cost time-consuming annotations. As a leveraging method, Deep Convolutional Generative Adversarial Networks (DCGANs) are proposed here to ameliorate this problem, they are trained on different-size partial subsets of one dataset and used to generate diverse and realistic mammographic lesions. The effect of adding these images is tested in an environment where a 1-to-10 imbalanced dataset of lesions and normal tissue is classified by a fully-convolutional neural network. We show that using the synthetic images in this environment outperforms the traditional augmentation method of flipping. A maximum of ~ 0.09 and ~ 0.013 improvement of F1 score and AUC, respectively, were reported by using GANs along with flipping augmentation compared to using the original images even with relatively-small dataset sizes. We show that DCGANs can be used for synthesizing photo-realistic mammographic mass patches with a considerable diversity measured using Frechet Inception Distance (FID).

Keywords: computer-aided detection, generative adversarial networks, data augmentation, breast cancer, deep learning, fully-convolutional neural networks, t-Stochastic Neighbor Embedding.

1. Introduction

1.1. Breast Cancer Detection

Cancerous breast cells have been the second deadliest disease in women globally coming after lung cancer. This disease was the most frequently diagnosed cancer in 154 countries and the first cause of cancer death in women in 100 countries in 2018 (Bray et al., 2018). In EU, breast cancer was the first cause of cancer death in 2014 for women, while for men it was lung cancer. Approximately, over 92000 women are anticipated to die because of breast cancer in 2019 with a similar number of deaths in 2014 (Malvezzi et al., 2019). Computer-aided detection (CADe) systems have shown that they can assist specialists in decision making although recent studies show that patient recalls have increased when using artificial intelligence as a second reader (Le et al., 2019). Moreover, CADe systems have been a good alternative for double reading to reduce failures resulted from mainly: visual search mistakes due to fatigue or other reasons, and mistakes in interpretation due to lack of decision-making experience for inexperienced interpreters (Bazzocchi et al., 2007). These systems can help reduce the diagnostic accuracy differences between radiologists caused by intra- and inter-observer variability (Elmore et al., 1994). Particularly, these systems can increase the sensitivity of less-experienced interpreters by increasing the detection rate by 10% (as a maximum) and reducing the time needed to detect the disease by one month in the best case. That said, the benefits observed in more-experienced readers is much smaller (Kohli and Jha, 2018). Additionally, CADe systems, represented currently by neural networks, require large amounts of annotated data when using supervised learning. Unsupervised learning is still under research where there is no need for completely-labelled datasets. However, a large number of experiments is usually needed

to teach the system making deep learning in general a limited-capability tool if the need to a large enough data is not met properly. Furthermore, publicly-available medical datasets are usually small and imbalanced due to concerns mainly related to privacy and the high costs needed to produce professional annotations by experts. To alleviate this problem of lack of data, different methods ranging from conventional data augmentation using affine transformations such as flipping or scaling, passing by sampling methods, to the more effective but complex way using Generative Adversarial Networks (GANs) (J. Goodfellow et al., 2014).

1.2. Generative Adversarial Networks

To use machine learning tools in CADe systems, a reasonable amount of medical data is needed to train the system on capturing abnormalities in input images that specialists try to detect. These abnormalities differ from one medical field to another, breast lesions and lung nodules, for instance. Public medical datasets usually suffer from unbalanced distribution of images between the classes under study. Target class images are commonly rare, for example in INbreast dataset, by Moreira et al. (2012), one fourth of the dataset contains breast lesions. Intrinsically, in a mammographic image, normal tissue patches largely outnumber lesion patches (the target concept), if there is any, when classifying with/without lesion images patch wise. When this kind of problem exists, the learning process becomes more difficult and sometimes might lead to a loss in generalisation (overfitting). To overcome this obstacle, scholars usually use different methods: oversampling the weak class (e.g. *SMOTE* and *ADASYN*), undersampling the strong class, or ensembling the weak class with subsets of the strong class to make multiple smaller balanced datasets (for instance *Easy Ensemble*, *BestCascade* and *NearMiss*) (He and Garcia, 2009). Most oversampling methods, if not all, in general, use either samples replication, interpolation, or extrapolation. By replication, the algorithm tries to push the population up by replicating some samples identically. Interpolation-based methods insert new samples that are derived from the neighbourhood by averaging the features, i.e., averaging two neighbouring samples belonging to the same class to find the midpoint sample. Finally, extrapolation methods, as image rotation, translation, and zooming, produce new samples that can increase the generalisation of the model by reducing (or removing) the contribution of some sorts of non-pertinent variance–differences that are unrelated to the discrimination process–in properties like image angle, center position, and size to decision making (Bowles et al., 2018). However, not all non-pertinent information is as easy to exclude from discriminative features as affine transformations, especially in medical imaging field where there is a lack of conventional augmentation tools to tackle all sources of non-pertinent variance. GANs, introduced in J. Goodfellow et al. (2014), made a revolution in the field of data synthesisation, where a network (called generator or G) learns the distribution of the input data implicitly by the aid of another network (called discriminator or D) which, in turns, tries to learn to distinguish real among fake (synthetically-generated) images and feed the result back to G to update the weights. In other words, G learns the mapping $Z \rightarrow X$, where Z is the latent space (noise) and X is the data distribution, while D learns the mapping $X \rightarrow [0, 1]$. These two networks learn simultaneously in order to get in the end a generator that can yield realistic and diverse images starting from a random input (latent vector). In theory, when the generator and discriminator become experts, G generates images that are classified as well as real images with a probability of 0.5, which is known as Nash equilibrium. To reach near this point, the learning process should converge in such a way that neither G nor D learns in a pace that is much higher than the other. Furthermore, GANs have the big advantage of being able to augment a wide range of variance sources providing that the dataset has enough examples. As an example, consider a breast mass detection problem where micro calcifications should not affect the decision, by applying traditional methods of augmentation it is hardly ever possible to add calcification to a mass-only lesion which can be done using a trained generator. Two main problems might come up when training GANs (Goodfellow, 2016):

- Mode collapse: this happens when the network generates images that are replications of one pattern with slight differences. In this case, G has a many-to-one mapping between the latent space and the output images. As a consequence, the diversity of the outputs will be low (low recall) while realism might be fine. In multi-class problems, two kinds of mode collapse (or equivalently mode dropping) might exist: intra-class and inter-class, where in the former kind, the generator synthesises images for which the per-class diversity is low, while, in the latter kind, G synthesises images from one (or a few) class(es) only, ignoring others.

- Oscillation: when the generator keeps generating different samples but with low realism (low precision) which are easy for D to reject. Meaning that the system never converges, this is commonly caused by imperfect tuning for the learning speed of G and D, where D learns quickly giving no time for G to improve. This results in G loss increasing early in the training process while D loss reaches low values.

In this paper, Deep Convolutional GAN (DCGAN) was selected due to training stability as presented in Radford et al. (2016). It was used to synthesise mammo-

graphic lesions to use them as data augmentation to support CADe for breast mass detection. The rest of this paper is organised as follows: section 2 describes in brief recent works on GANs in medical imaging. Materials are explained in section 3, while methods are presented in section 4. Sections 5, 6, and 7 include the results, discussion and conclusions, respectively. The contributions of this work are as follows:

1. We show that DCGANs are able to generate images of 128 × 128 pixels of realistic and diverse mammographic mass and calcification lesions evaluated quantitatively using Frechet Inception Distance.
2. We tested the DCGANs performance after being trained and we show that it provides remarkable improvements when used to augment an imbalanced dataset.
3. We analysed the effect of adding the synthesized images to an imbalanced dataset as a function of training set size in a classification problem.
4. We propose one framework (Figure 7) under which all previous points can be tested combined using 3-fold cross validation.
5. We show that the generated images belong to the real images' distribution by visualizing the t-Stochastic Neighbor Embedding (t-SNE) of both real and fake images.
6. We made the trained generators publicly available, along with the code, to the scientific community to generate patches of breast masses.

2. State of the art

Korkinof et al. (2018) used Progressive GANs to generate 1280 × 1024 full mammogram images that show breast anatomy with acceptable amount of fine details using the multi-stage adversarial learning introduced in (Karras et al., 2018). In Salehinejad et al. (2018), 5-class chest pathology X-ray 256×256 images were generated using the well-known DCGAN architecture by Radford et al. (2016). They evaluated the effect of including the synthesized images by measuring the balanced test accuracy using three models, namely: real imbalanced dataset (DS1), real balanced dataset (DS2) with 2K images from each class, and balanced real + synthesized images (DS3) with approximately 30K images from each class. The results clearly showed that including synthetic images boosted the performance of the model significantly with average accuracies DS1: 70.87, DS2: 58.90, DS3: 92.1. Conditional infilling for mammogram lesions was presented in Wu et al. (2018), where the authors filled a masked region in a patch with a multi-stage training approach (similar to resolution pyramids). They proved that starting with small generated images then enlarging them gradually gave high resolution images that were useful to augment the unbalanced dataset and get a higher AUC value. They used different kinds of loss, where to assure realism they used feature loss which is the average of squared differences between the pretrained-VGG-19 feature maps from real and fake images, but they used boundary loss to get smooth edges between the generated lesion and the infilled component by minimising the difference at the boundary of the lesion. To evaluate the outcome of the generator objectively, ResNet 50 was used as a classifier to show performance improvement, ciGANs model combined with traditional augmentation was reported to have a +0.014 AUC more than the baseline model (without augmentation) and +0.009 than traditional augmentation. Frid-Adar et al. (2018) used GANs to generate 2D liver lesions by training a DCGAN on 182 images belonging to three classes with conventional augmentation applied on the input. Thereafter, they generated images by the trained generator and used these images as augmentation over the conventional methods of rotation, scaling, and translation. They showed that GANs improved sensitivity and specificity by 7% and 4%, respectively, with respect to using traditional augmentation methods only. Additionally, they showed that using t-distributed Stochastic Neighbour Embedding (t-SNE) tool, GANs can provide more diverse features than traditional augmentation. However, they did not investigate the effect of changing the size of training set on GANs images quality, and consequently, on the DCGAN-augmented classification problem. They presented that two observers were able to achieve approximately 62% and 58% accuracy in differentiating real from fake images, but they did not show any Frechet Inception Distance (FID) or Inception score (IS) to evaluate the realism and diversity of their synthesized images objectively. They found out that adding more synthetic images beyond some limit did not improve the classification performance any more and they analysed on a small scale the effect of adding a few more real images. Bowles et al. (2018) used DCGANs to generate synthetic segmented Computed Tomography (CT) and Magnetic Resonance (MR) brain images to enhance the performance of segmentation networks. They included an interesting experiment where they studied the effect of applying conventional augmentation (rotation, flipping, scaling) on DCGAN-generated images and they found out that the traditionally-augmented GANs images could improve the performance more than the sum of GAN and augmentation improvements when trained separately. Another important point they highlighted was that GANs do not impose any negative impact on the classification performance when trained on limited datasets, on the contrary, when the GAN was trained on a relatively large dataset, it introduced a decay in the overall segmentation performance. Douzas and Bacao (2018) exhaustively compared conditional GANs (cGANs) by Mirza and Osindero (2014) with SMOTE

Table 1: Dataset Annotations given in Excel files.

Image Information	Patient ID Study ID Series ID Image ID
Lesion Information	Lesion ID x1, y1 x2, y2 Lesion status Lesion type

by Chawla et al. (2002) and its variations of oversampling methods using 71 datasets with different sizes and imbalance ratios. They used 5 different classifiers; Support Vector Machines, Decision Trees, Logistic Regression, Gradient Boosting machines, K-Nearest Neighbours; and three metrics: F score, G mean, and Area Under the ROC Curve (AUC). In conclusion, they reported that cGANs statistically outperformed other methods and had the highest mean rank (closer to one) using all datasets , classifiers, and metrics. However, they did not include any deep learning method as a classifier and no qualitative evaluation was mentioned.

3. Materials

The dataset used in this work was OPTIMAM Halling-Brown et al. (2014) which has around 80,000 processed and unprocessed images extracted from the National Breast Screening System (NBSS). This dataset has expert annotations linked to images via exhaustive Excel files that have all the information required to identify the image and any clinical observation. Table 1 shows some column headers of the Excel files. Image information fields link the image to a patient, a study (where some patients have more than one study), and a series. Lesion coordinates (x1, y1, x2, y2) are given in pixels, lesion status can be one of: Breast Imaging Reporting and Data Systems (BIRADS) levels: B1, B2, B3, B4, or B5, where B1 categorizes the finding as negative while B5 is for highly suspicious of malignancy (Orel et al., 1999). Lesion type can be one of: mass, calcification, focal asymmetry, architectural distortion or a combination of them. Table 2 shows more columns that were used later on to filter the dataset. Images included in this dataset were acquired using modalities made by different manufacturers: Philips, General Electric, Hologic, Faxitron X-ray Corporation, Lorad, Siemens, or Bioptics Inc. Model name is the model of the device used for acquisition (examples are Selenia, Bio Vision, and L30 Philips). X-ray tube current is the estimated value of the current used to acquire the image (ranges from 1 to 1500 mA) with a specific magnification factor (ranges from 1 to 2.15). Additionally, presentation states whether the image was processed from origin or not. In summary, the dataset was heterogeneous combining images from different manufacturers and modalities which resulted in a wide spectrum of distributions. As it is known in classification problems using deep learning tools, training images should come from similar distributions so the network can learn the general pattern. Figure 1 shows four different images from the dataset. Images from the left column have the same properties (modality, manufacturer, settings) but still (c) has some measurements that cannot be included in training the GAN, resulting in filtering these properties due to the difficulty in distinguishing between cases like (a) and (c). Case (d) has a distribution that is not aligned with other images where the background is white and dense tissues are represented by dark intensities. Case (b) is a sample from the set of properties selected where the contrast is relatively better than other cases.

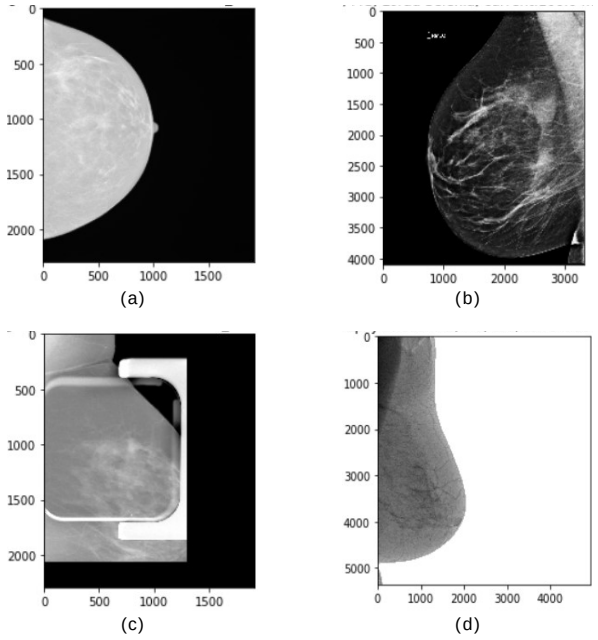


Figure 1: Some different samples that show the importance of filtering the dataset. (a) is a CC-view of a mammogram by GE Medical Systems with current 62 mA and magnification factor 1.0 , (b) an MLO view by Hologic Selenia, current 100 mA and magnification factor 1.0, (c) an MLO by GE Medical Systems Senographe Essential with current 62 mA and magnification factor 1.0 (notice the magnification view), (d) an MLO by Philips Digital Mammography Sweden L30 with current 180.0 mA and magnification factor 1.0304.

3.1. Image Selection Criteria

In order to properly train a neural network, the input images should have similar distributions. To satisfy this requirement, a filtering technique was applied on the dataset using the annotation files. Exhaustive experiments were conducted to show images belonging to different sets of configurations. It turned out that images

Table 2: Acquisition settings criteria.

Criterion	Value
Manufacturer	Hologic, Inc.
model name	Lorad Selenia
X-ray tube current	100
Magnification Factor	1.0
For presentation	True

with the characteristics shown in Table 2 had similar intensity distributions, so they were selected for extracting patches. The idea is first to select one manufacturer and one model which are Hologic and Lorad Selenia from where more than half the dataset came. By doing this, all selected images have experienced the same processing. Second, to avoid images with special magnified projections (see Figure 1 (c)), the current was fixed to 100 mA and magnification factor to 1. Lastly, only the processed images were used. After this filtering, 14,549 lesion-free images in addition to 5267 with-lesion mammograms were selected including Craniocaudal (CC) and mediolateral Oblique (MLO) views from right and left breasts. This data belonged to 3701 patients (some patients had more than one study and sometimes more than one lesion per image).

3.2. Breast Mask Generation

OPTIMAM mammographic images come with no breast masks, however, there was a need to extract patches background free. To meet this need, a simple thresholding algorithm ($I > 0$; I is a grayscale image) was applied on the filtered dataset. In other words, if a pixel has a non-zero intensity, it will be considered part of the breast, see lines 1-4 in Algorithm 1. As an example of the mask, see Figure 2. All mask images were saved with meaningful names by adding the extension _msk to the original image name. In order to make the process of finding the corresponding pair (image,mask) straightforward, the original folder architecture (*batch* → *patient* → *study*) was preserved.

3.3. Lesion Groundtruth Localization

To generate lesion patches, a groundtruth image was needed as a reference. To generate these images, a simple process was followed (see Figure 2). First, the lesion coordinates (x1, y1, x2, y2) are extracted from the the Excel file. Second, an empty image with the same size of the mammogram image is created then the area between (x1,y1) and (x2,y2), including endpoints, is filled with the value 255 (not 1 for visualization issues). Third and last, the groundtruth image with the corresponding image name adding the extension _gt to the end and keeping the original folder architecture (*batch* → *patient* → *study*) is saved, see lines 5-9 in Algorithm 1.

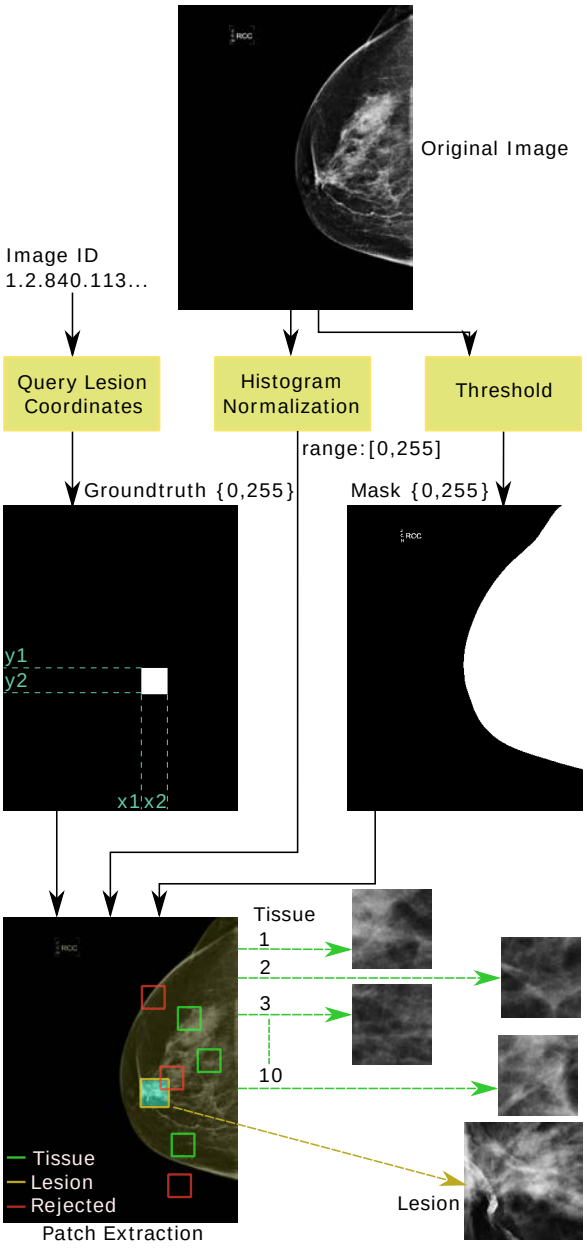


Figure 2: Data preparation overview. From top to bottom and left to right: original image, query lesion coordinates represents the process of reading the lesion top-right and bottom-left coordinates from the database, histogram normalization represents the process of transforming the intensity distribution to the range [0, 255], non-zero thresholding, lesion groundtruth, breast mask, overlayed images (yellow overlay: breast mask, cyan overlay: lesion groundtruth, grayscale: histogram-normalized image), tissue patches (green rectangles) and lesion patch (yellow rectangle). The red rectangles represent rejected patches, where normal tissue conditions are violated (the top red rectangle has some background, the middle one is located partially inside the lesion, while, the lowest one has complete background pixels).

3.4. Image Preprocessing and Patch Extraction

Image preprocessing and patch extraction steps are summarized in Figure 2. Using the filtering criteria mentioned in section 3.1, images which do not meet the inclusion criteria were not included. After filtering, we read one image (top block in the figure) and create the corresponding mask image (see section 3.2). If this image contains a lesion, the lesion groundtruth image is created as described in section 3.3, otherwise an empty groundtruth image is created (see the left branch in the figure). Histogram normalization was applied on all filtered images to assure similar intensities range [0, 255] and data type (unsigned integer on 8 bits). All three outputs (normalized image + mask + groundtruth) were used to extract the patches as depicted in Figure 2 bottom part where only three green rectangles are shown for the sake of a simple figure. In practice, ten random normal-tissue patches of 128 × 128 pixels were extracted, in addition to full-lesion patch(es) if any lesion exists. In Algorithm 1, the first three lines read an image from the filtered dataset after applying histogram stretching. Then, the corresponding mask and groundtruth (see sections 3.2 and 3.3) are created and the lesion patch is saved with dimensions that might be different from one lesion to another. Starting from line 12 in the algorithm, patch extraction process includes extracting 100 random patches and verifying if they belong to breast region with white mask (see the first part of the condition at line 15). Normal tissue patches are lesion free as indicated in the second term of line 15 in the algorithm. The algorithm keeps extracting patches until it reaches 10 valid patches or 5 iterations before stopping. In this case, for every image, a maximum of 10 normal tissue (referred to as 'Tis') patches and a number of lesion patches ('Les'), which is related to the number of lesions contained in the image, are extracted.

3.5. Patches Post-Processing

In addition to the previously-described preprocessing steps, some of 'Tis' patches had a very narrow histogram. Those patches did not carry enough intensity variations to resemble normal tissue patches. A simple post processing algorithm was applied in which the number of unique intensity values for each patch was calculated. Patches with less than 30 different intensity values were removed from the patch dataset. After all, 5351 lesion (classes include mass, calcification, focal asymmetry, and architectural distortion) and 147,951 normal tissue patches, extracted from all filtered mammograms regardless having a lesion or not, were saved for training the GAN and the classifier.

4. Methods

4.1. The Generator

As mentioned before in the introduction, DCGAN by Radford et al. (2016) was used with some modifications in this work. The architecture of G is shown in Figure 3. The aim of the generator is to learn the mapping between the latent space (the normal distribution in this case) and the space of mammographic lesions in a sense that it can transform a vector from the latent space to a lesion image that can fool the discriminator. Figure 3 shows that the generator (with green color referring to G in this work) had six layers (it was 5 in the original paper ending with 64×64 output). The first layer projects the latent vector and reshapes it to the first cube shown. Internally, it is a dense layer followed by reshape. Tconv2d refers to Transpose Convolution 2D with kernel size 4, stride 2 and one pixel padding. In this implementation, no max pools nor dense layers were used as suggested in Radford et al. (2016). The activation function used was LeakyRelu with negative slope 0.2 and batch normalization on all layers except the last one where the activation function was hyperbolic tangent (Tanh).

4.2. The Discriminator

The discriminator task is to distinguish between real and fake lesion images outputting realism probability (0 means definitely fake, 1 means definitely real). Figure 4 shows the architecture of the discriminator where it accepts an image, it resizes it to 128×128, and normalizes its intensity to the range [-1, 1]. The six layers (five in the original paper) are similar to the generator's ones but the opposite direction. Convolution2d layers were activated by LeakyRelu with negative slope 0.2. 2D batch normalization was used in all layers except the first and last ones. Stride 2 was used to downscale the size until layer 6 where stride was 1. The kernel size 6×6 was used for all layers with padding two (except for last one 4×4 and 0 padding). The activation function for the last layer was sigmoid to output a probability between 0 and 1.

4.3. DCGAN Training

As mentioned in Lucic et al. (2018), GANs losses do not matter as hyperparameter tuning and the availability of computational resources. The loss functions used to train this DCGAN were the ones recommended in J. Goodfellow et al. (2014), see equation (1) for discriminator loss ($J^{(D)}$) and (2) for generator one ($J^{(G)}$). To give a brief explanation of these loss functions, the discriminator loss is aiming to provide values as close to 1 as possible for real inputs (maximize $log(x)$), while, giving as close to 0 as possible for fake inputs (maximize $log(1 - D(G(z)))$). For G loss, this is the modified version proposed in J. Goodfellow et al. (2014), where the generator tries to fool the discriminator to get as close to 1 as possible by generating images that D gives high realism probabilities, this loss is referred to as Non-Saturating loss (NS loss). The convergence occurs when the discriminator cannot actually distinguish

Algorithm 1 Patch Extraction

read one image from the filtered preprocessed dataset, I
$H, W = size(I)$
$mask = zeros(H, W)$
$mask[I > 0] = 255$
$GT = zeros(H, W)$
if hasLesion **then**
    fetch lesion coordinates (x1, y1, x2, y2)
    $GT[y1 : y2, x1 : x2] = 255$
    Save I[y1:y2, x1:x2]
end if
$Count = 0$
while $Count < 10$ **and** $max\_iter < 5$ **do**
    extract 100 random patches $(p_0, p_1, \ldots, p_{99})$

$$p = extract\_patches2d(I, num = 100, size = (128, 128))$$

    for all p_i **do**
        if $\sum mask[p_i \cap mask] == 255 \times 128^2$ *and* $\sum GT[GT \cap p_i] == 0$ **then**
            Save p
            $Count++$
        end if
    end for
    max_iter++
end while

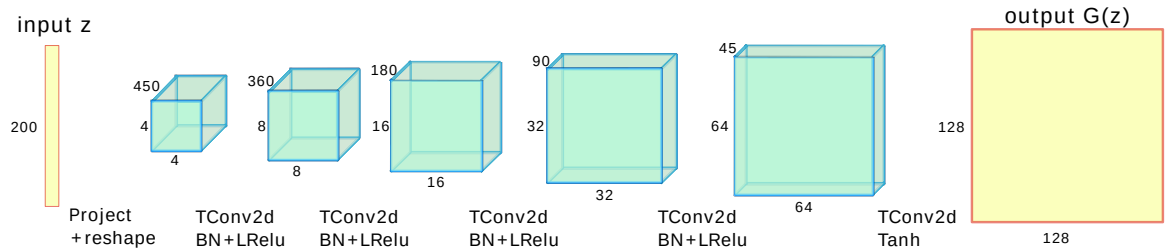


Figure 3: Generator architecture, the input belongs to the normal distribution with 0 mean and 1 standard deviation, TConv2d represents a transpose convolution 2D (kernel size 4, padding 1, stride 2 except for the first one where stride=1, padding=0), BN stands for 2D batch normalization, LRelu means leakyRelu with a 0.2 negative slope.

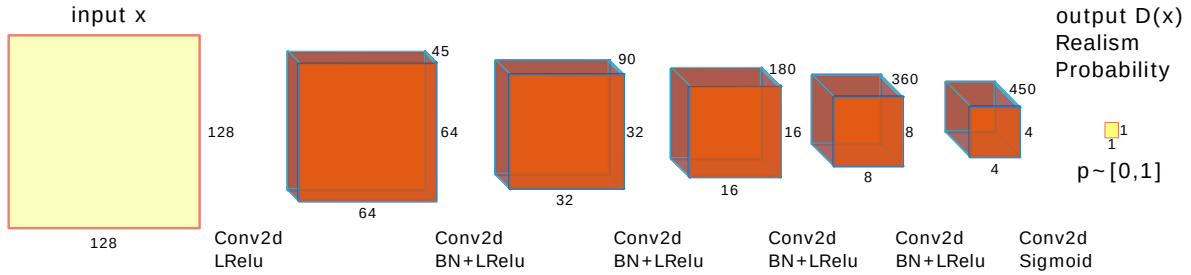


Figure 4: Discriminator architecture, Conv2d represents a convolution 2D layer (kernel= 6, stride= 2, padding= 2, except for the last one where kernel=4, stride=1, padding= 0), BN stands for 2D batch normalization, LRelu means leakyRelu with a 0.2 negative slope.

real among fake cases where the ideal case is to have 0.5 on the output of D for both real and fake inputs (Nash equilibrium), meaning that the distributions (P_x, P_{gen}) are completely matched and there is no possibility to find the boundary for the classifier.

$$J^{(D)} = -E_{x \in P_x,\, z \in P_z}[\log(x) + \log(1 - D(G(z)))] \quad (1)$$

$$J^{(G)} = -E_{z \in P_z}[\, log(D(G(z)))\,] \quad (2)$$

In Equations (1,2), E refers to averaging over training examples, P_x, P_z refers to training images and noise distributions, respectively, z is the random vector input to G, and $G(z)$ is the synthetic output of G. As mentioned in J. Goodfellow et al. (2014), this G loss is preferred to $log(1 - D(G(z)))$ because it has higher gradients at the beginning of the training process which makes G learn faster, see Figure 5. The optimizer used to train both G and D was Adam by Kingma and Ba (2014) with $\beta_1 = 0$, $\beta_2 = 0.99$ and learning rates $4e^{-4}$, $2e^{-4}$ for G and D respectively. Learning rates were exponentially decreased by a factor of 0.99 every 10 epochs for G, and 8 epochs for D. The batch size was 64 and the model was trained for 1000 epochs. Figure 6 shows the training procedure step by step, where the dense arrows refer to real-image related processes, while the dashed ones refer to synthesized-image related processes. Every training iteration, a batch of random latent vectors are generated from the normal distribution with zero mean and unit standard deviation ($z \in P_z$; $P_z = \mathcal{N}(0, 1)$), see step 1 in the figure. This pure-noise batch is to be first normalized to the range $[-1, 1]$ then forwarded through G to generate a batch of fake images (G(z)), see step two in the figure. These fake images are first normalized to the range $[0, 1]$ then forwarded through D to get realism probabilities, see step three with dashed arrows. An equal-size batch of real images is normalized and forwarded through D to learn the boundary between real and fake lesion spaces, see step three dense arrow. In step four, equation (1) is used to calculate the loss for the discriminator, then backpropagation is done to update D parameters, see step five. Equation (2) is used to calculate G loss in step six. Then, backpropagation is done to update G parameters, see step seven. To complete one epoch, this process, from step 1 until seven, is repeated until all the real images are covered.

4.3.1. Training Techniques used

Training GANs is a precise process that should be driven carefully to avoid divergence problems (see section 1). In this work, different work-arounds have been used to overcome common problems, such as getting similar lesions all having the same shape with slight differences or even getting unrealistic lesions (see the early stages in Figure 14 in annex 9.2). As mentioned in Salimans et al. (2016), one-sided label smoothing was a useful technique in which over-confidence problems were resolved. Every epoch, a value in the range [0.7, 1] is

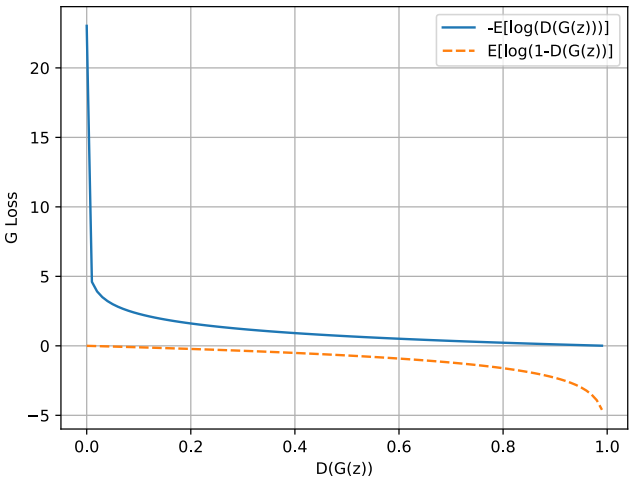


Figure 5: Generator loss comparison between original minmax loss (orange dashes) and the non-saturating loss function (the blue solid) in (J. Goodfellow et al., 2014). $D(G(z))$ represents the discriminator output for the generated images and E represents averaging over the number of generated images.

picked to be the real label for training D and G which helped to force the discriminator to keep learning so the gradients never diminish which, as a result, pushes G to keep enhancing the output results. This does not affect the accuracy (the real label is still higher than 0.5). Conventional data augmentation, horizontal and vertical flipping, was used in which the original dataset size does not change as the flipping happens on the fly. This helps increase the diversity of the generated images. One of the critical issues that were faced during training was the checkerboard effect in which a rough grid shows up in the synthesized images, which obviously reduces the realism. The explanation of the problem was that this artifact was in the blind spot of D (because D and G kernels were completely aligned before changing) so it did not contribute enough to the loss function. The solution was inspired by a talk of Goodfellow (2016) [1] where it was suggested to use different kernel sizes between G and D, so a larger kernel (6×6 instead of 4×4) was used for D which made that artifact more visible to D and it could penalize G for it. The results were significantly improved with a noticeable increase in diversity and realism as well. Other techniques, namely: spectral normalization as in Miyato et al. (2018), layer normalization as in Lei Ba et al. (2016), pixel shuffle to resolve checkerboard effect as in Shi et al. (2016), and label flipping were used but no significant effect on the results was observed.

4.4. DCGAN Evaluation

To evaluate the generated images by the DCGAN, different tools were utilized as follows.

[1] The video is available at: https://channel9.msdn.com

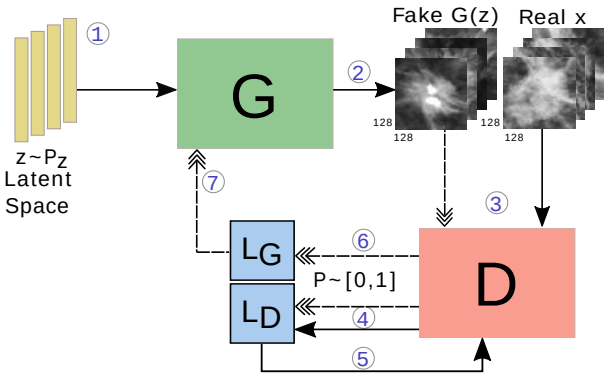


Figure 6: The top view of training DCGAN, the input belongs to the normal distribution P_z with mean 0 and standard deviation 1. Dotted arrows refer to fake-input related values. Steps from one to seven are: generate a noise batch, forward through G to generate a fake batch, forward the real and fake batches through D, calculate L_D, update D, calculate L_G, and update G, in order.

4.4.1. Training Phase

Frechet Inception Distance (FID), proposed by Heusel et al. (2017), was calculated to reflect the performance of the generator during the process of training the DCGAN. The equation for calculating FID is:

$$FID(R,G) = \|\mu_R - \mu_G\|_2^2 + Tr(\Sigma_R + \Sigma_G - 2\sqrt{\Sigma_R\Sigma_G}) \quad (3)$$

where R, and S are the real and synthesized images folders, respectively. μ_R is the mean of feature maps of Inception-v3 by Szegedy et al. (2016) for the real folder, μ_S is the mean of Inception-v3 feature vectors for the synthetic folder. Σ_R is the covariance matrix of Inception-v3 output feature vectors for the real folder, Σ_S is the covariance matrix of Inception-v3 output feature vectors for the synthetic folder. Tr is the trace process of adding the elements of the main diagonal.

The idea is to use Inception-v3 pretrained on ImageNet as a feature descriptor for real and fake images (using 2048 activation units), then to calculate the difference between the means of the two folders as well as the the second term in equation(3) [2]. Within GANs users, Inception score by Salimans et al. (2016) is one of the frequently used evaluation metrics, however, this metric has to be computed over large enough generated/real images (50K as mentioned in the paper) which is ten times larger than the number of positive examples in this work. Additionally, Heusel et al. (2017) showed that FID is more robust against noise and more consistent than Inception score, in other words, the more similar the generated images to real ones, the lower the FID. Other works proved mathematically that Inception Score worked well on ImageNet but it is not guaranteed to be working as well on other, especially smaller, datasets (Barratt and Sharma, 2018). Additionally, IS captures precision and inter-class diversity while it fails to capture intra-class diversity which are all captured by FID (Lucic et al., 2018). Due to all preceding, FID was preferred as a guideline during training and sometimes as a model-saving criterion, see Figure 8 for an example of FID progress during DCGAN training. Regarding overfitting, neither FID nor IS can capture because they are intrinsically optimal when the generated images match the training ones.

[2] Implementation in Pytorch was adapted from `https://github.com/mseitzer/pytorch-fid`

4.4.2. Testing Phase

In order to evaluate the trained generator, an augmentation environment is used where an imbalanced dataset of lesions (positive minority class) and normal tissue (negative majority class) is to be classified by a fully-convolutional neural network. In this setting, the classifier has almost the same architecture as the DCGAN discriminator with slight differences (less filters due to a smaller dataset) using $\{9, 18, 36, 72, 90\}$ as number of channels, from first to last layer respectively (see Figure 4). Additionally, 5% weight decay was used as a regularizer. Furthermore, the distribution of the generated images was compared to the real ones' in the two-dimensional space of t-SNE (van der Maaten and Hinton, 2008).

4.5. DCGAN for Lesion Simulation

In this work, the DCGAN was trained to generate mammographic lesions that look like real ones (visually indistinguishable) using 4536 mass and calcification lesions. Other lesion classes of architectural distortion and focal asymmetry were not used because in such classes the lesion existence in one breast location is captured when the two breasts look asymmetric at this location, (Samardar et al., 2002). This simultaneous observation of both breasts was infeasible for the classifier. Horizontal then vertical random online flipping was used as augmentation. As the complete dataset had mammographic mass and calcification lesions (sometimes in the same patch), the GAN was trained to generate mass, calcification, or both in the same patch. These settings have the advantage of a relatively-large dataset where the generator can see a wide spectrum of cases to capture the distribution. The application of this mode is to train radiologists/observers or other specialists on different tasks related to lesion detection and annotation on unseen images with a considerable quality that are hard to distinguish from real patches. This environment has a limitation that it gives just a small part of the big picture, i.e. a patch out of the complete x-ray image. Consequently, it might be hard to detect lesions related to architecture asymmetry and architectural distortions where the corresponding patch of the second breast is

needed to compare to. The resultant generators can be used as an augmentation tool in cases where there is no need to separate calcification from masses (considered as one class), however, this application was not studied here in favour of mass class augmentation where the DCGAN output has a predetermined class which can make the process of evaluating the augmentation effect independent from the class of the generated images, see section 4.6 for more details. In this scenario, the GAN was trained using the hyperparameters mentioned in section 4.3. It is worth mentioning that long training as well as mismatched sizes for G and D kernels were useful for increasing images quality (to get rid of checkerboard effect generated by transpose convolution layers) and diversity, but that should be accompanied by a fine-tuned learning rate decay (we used here as mentioned before $4e^{-4}$ and $2e^{-4}$ with 1% decay every 10 and 8 epochs for G and D respectively). It is common, as observed in Figure 14, that D wins at the end of the game by obtaining a smaller loss with respect to G, however, this should not be very early in the training, otherwise, the generator will find difficulty learning from small gradients, which ends up with the GAN diverging. Figure 12 in annex 9.1 includes two 8×8 batches, showing real and generated images.

4.6. Mass Lesion Augmentation Using Different Training Sizes

The aim of this method is to analyse the following points:

- The effect of increasing the size of the training set of the positive class, by adding real images ,on the performance of a classifier keeping the same imbalance ratio (IR equals to 1:10).
- How the random online augmentation (horizontal flipping followed by vertical flipping with a probability of 0.5 for each) affects the classification performance in this unbalanced environment.
- The change in classification performance after adding the DCGAN-generated images to the dataset keeping fixed the augmentation ratio AF = 1.5 and IR= 10 again as a function of the training size.

To clarify all previous points, a framework is proposed (see Figure 7) which was inspired by the works of Frid-Adar et al. (2018), Bowles et al. (2018), and Douzas and Bacao (2018), where in in the latter they trained the GAN on the training set of the classifier to avoid generating images that might have features similar to the test/validation images'. We combine the idea of studying the effect of changing the number of the images used to train the GAN, as well as applying conventional augmentation methods on the generated images. This was examined on a small scale in Frid-Adar et al. (2018) due to lack of data, while in this work we had the advantage of using a larger dataset. The dataset used to train the DCGAN was a subset of the dataset described in section 3.4, where 2215 mass lesion patches (positive class) were selected, including benign and malignant cases. After extracting the test set (33.3%), the remaining part was divided into training and validation (60%, 6.6%, respectively), and finally the training part was divided into six overlapping smaller sets: $\{P_k; k \in \{100, 250, 500, 750, 1000, 1300\}\}$, where the subscripts refer to the size of the subset. All these subsets were picked randomly with a fixed seed for the random generator so that each set is contained in the next larger one. For instance, $P_{100} \subset P_{250} \subset P_{500}$, see the dataset and sampler part of Figure 7. Regarding the negative class (normal tissue patches), a similar procedure was applied on a 22K subset selected randomly out of the 147K complete normal tissue dataset to have an IR of 1:10. Six overlapping negative subsets with the size ten times the positive class were created to use later on in classification, namely $\{N_{1000}, N_{2500}, N_{5000}, N_{7500}, N_{10000}, N_{13000}\}$. For training the DCGAN, one positive set P_k was used at a time, and due to the use of relatively smaller datasets than the one used in section 4.5, a few hyperparameters were changed: horizontal then vertical flipping was applied as before, in addition to jittering the brightness and contrast by a random amount picked from the range [−5, +5] every iteration. Furthermore, the DCGAN was trained for 1000 epochs to give the generator enough time to learn the distribution. These settings were fixed for any k. After training six DCGANs independently, six generators $\{G_{100}, G_{250}, G_{500}, G_{750}, G_{1000}, G_{1300}\}$ were ready to generate synthetic mammographic patches (size 128×128), see the top right part of the figure. Thereafter, four classification modes were investigated (see the middle part of Figure 7), namely:

- *ORG*: in this mode, the input for the classifier is P_k as positive images plus N_k as negative. The aim of this mode was to see how changing the positive class size affects the overall classification performance keeping IR 10 for all cases.
- *Aug ORG*: as the name suggests, augmented original images were used as input to the classifier. By augmentation here we mean random horizontal then vertical flipping ending up with one of the following cases:
  - Only horizontal flipping.
  - Only vertical flipping.
  - Both horizontal and vertical flipping.
  - No flipping.

  No intensity or rotation/translation augmentation were introduced here to preserve the content from

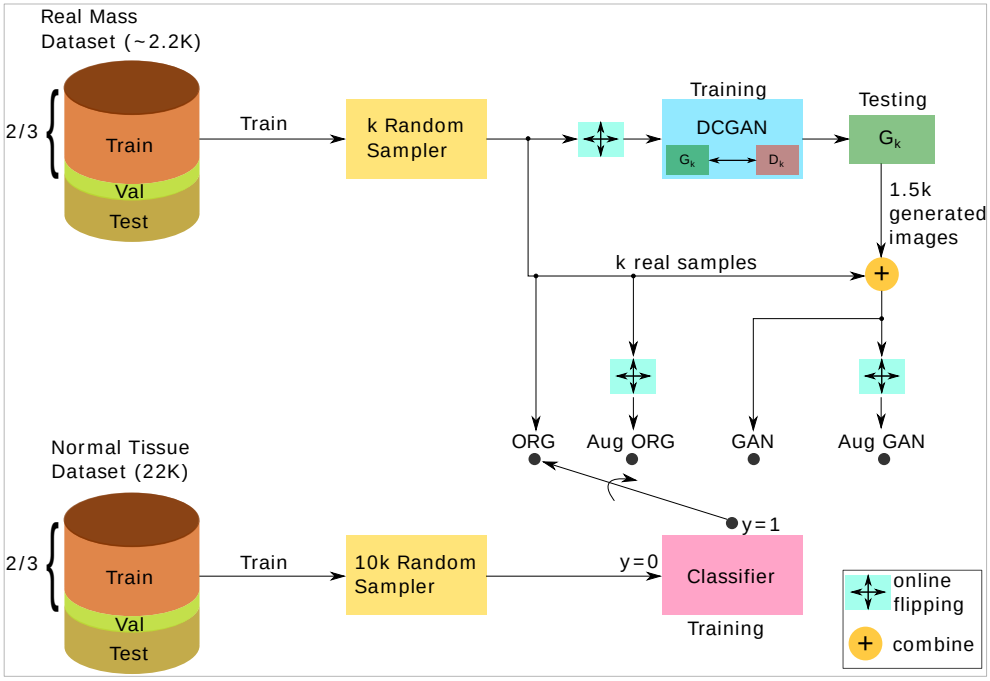


Figure 7: The proposed framework for evaluating the DCGAN when used in data augmentation for supporting the minority class in an unbalanced dataset.

including any padding or interpolation. The aim of this mode is to study the effect of conventional augmentation on the unbalanced classification problem as a function of the the positive class size , keeping the imbalance ratio fixed to 10.

- *GAN*: the input to the classifier in this mode was k real lesion + $1.5 \times k$ synthetic images generated by G_k as the positive class, and $10 \times k$ normal tissue patches as the negative class. The aim of this mode was to analyse the effect of combining the synthetically-generated images with the real images to support the under-represented positive class in the classification problem. $1.5 \times k$ was selected to give G_k the chance to reflect the learned distribution with a reasonable diversity. The aim of this mode was to inspect the effect of using multiple DCGANs trained on datasets with different sizes on the classification problem.

- *Aug GAN*: in this mode, the $1.5 \times k$ generated images as well as the real ones were augmented on the fly by random flipping (same as *Aug ORG*) to extend more the distribution of the input images. The aim was to see whether flipping the synthetic images would add any valuable features to the classifier.

Flipping is considered as an extrapolation method as opposed to GANs which are considered as an interpolation method (Bowles et al., 2018). To use flipping only was inspired by the works of Kamnitsas et al. (2017) and Wu et al. (2018), where they preferred to use reflection only to preserve the architecture without using any intensity perturbations. This is particular for medical images where other affine transformations can change some discriminative features in the patch, for instance rotation might have introduced padding pixels while zooming can change the lesion size which may have an impact on decision making. The classifier used here is depicted in Figure 7 the bottom right part. It has the same network architecture as the DCGAN discriminator apart from number of channels, see section 4.4.2. It was trained with the same parameters but for fewer epochs and binary cross entropy as a loss function (instead of the adversarial loss in DCGAN equation (1)), where for all modes, 20 epochs were enough to reach almost 100% training accuracy. Knowing that the dataset is imbalanced, using accuracy might be misleading and might give very high values even for a naive classifier that outputs the negative label always. As a result, F1 score was proposed to be used as a metric which gives equal importance to precision and recall, see equation (4).

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

To avoid an overfit classifier, for each mode, the model with the best validation F1 score was saved for testing phase. As can be seen in Figure 7, the test and validation sets were fixed for all k. 3-fold cross validation was used to acquire reliable results.

5. Results

5.1. Lesion Simulation

After training the DCGAN for 1000 epochs with a decaying learning rate, D has been trained to discern real from generated lesions, and in the same time, G has learned how to generate lesions that look like real ones by getting skilled more and more as the training progresses. In line with FID concept explained before (see section 4.4.1), Figure 8 shows how the value of FID changes during training where it starts with a value around 120 and drops drastically until it reaches a plateau around 20 where generated and real images look similar for Inception-v3 network. The orange line representing average FID is used to show the trend where the more the DCGAN is trained the lower the average FID is until convergence. This can be explained by the fact that at the beginning of the training process, the quality of the generated images is far from the real ones' which makes the discrimination task easy for D, so it gives very low realism probabilities for G outputs, consequently G learns quickly, see Figure 5 to see the fast-moving loss function at the beginning (the blue line causes larger gradients and faster learning). In testing phase, the trained G is capable of generating any number of images by forwarding the same number of random vectors. A batch of 64 generated images is shown along with the same number of real ones in annex 9.1.

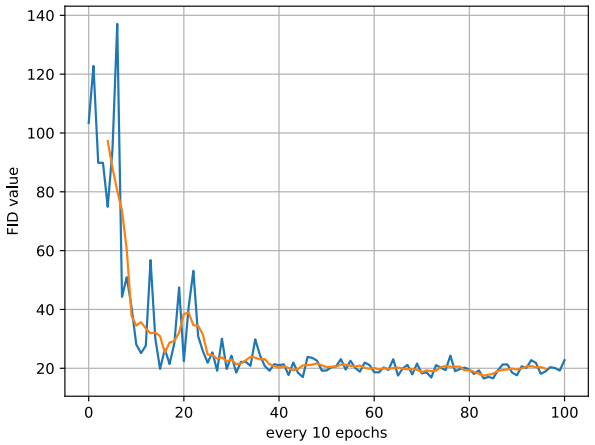


Figure 8: FID progress along the training process. The orange line represents the average FID over five neighbouring points where the blue values were recorded every 10 epochs.

5.2. Mass Lesion Augmentation Using Different Training Sizes

The results for methods in section 4.6 are presented here where the effect of adding different numbers of generated mass lesions as well as real ones was analysed in an imbalanced environment with IR = 10 over three cross validation folds. By looking at Figure 9, the blue line representing mode *ORG* is behaving in a way that shows that adding more real images helps the classifier

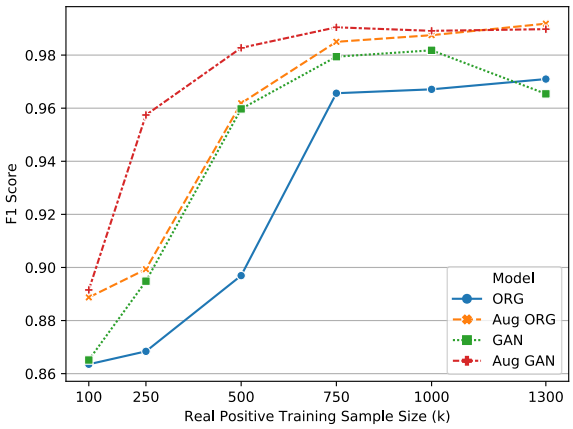


Figure 9: Examining F1 score as a function of the real minority training set size when adding 150% generated images and keeping the same imbalance ratio 1:10. The horizontal axis represents the size of the training set of the positive class (mass lesions). ORG stands for the original dataset without any kind of augmentation, Aug ORG represents using the online random flipping on the original dataset, GAN means combining the original dataset with the generated images without augmentation, Aug GAN refers to online flipping applied on real (positive and negative) and DCGAN-generated images.

to perform better regarding F1 score where it keeps improving until k=750 where it saturates after improving F1 score by approximately 0.1 compared to when using 100 images.

The green line represents the F1 score when adding synthetic images to the real ones (*GAN* mode). The amount of added images differs from one case to another but always using $1.5 \times k$ as augmentation factor. Compared to the blue line, the green one shows faster improvements which shows that the generator has learned to unlock unseen images in the real distribution which help the classifier to distinguish lesions among normal tissue. At 100, as the plot shows, the improvements were fairly existing which is due to lack of enough samples for the DCGAN to learn the distribution of the real data. As this amount increases to 250, the improvement over ORG increases drastically pointing at a better-performing G. This improvement continued until 1000 where the classifier was no more starving for data. Surprisingly, at 1300, the DCGAN could generate samples visually similar to the real ones (see annex 9.4 for an illustration of the effect of increasing the training set size on the generated images by DCGAN). However, this had a negative impact on the classification problem which might be due to overfitting. Additionally, the amount generated at 1300 ($1.5 \times 1300 = 1950$) is the largest among all experiments which might have caused a drop in diversity. Moving on to the orange line representing *Aug ORG* mode where online horizontal and vertical flipping with probability 0.5 was applied on every batch. Flipping was clearly outperforming *GAN* due to lack of training images for the DCGAN as opposed

Table 3: Area Uner the ROC Curve (AUC) for different modes and training sizes (k), the bold-faced values are the highest.

Mode	Training Size					
	100	250	500	750	1000	1300
ORG	0.9836	0.9848	0.9896	0.999	0.9989	0.9989
GAN	0.9843	0.9902	0.9984	0.9997	0.9993	0.9987
Aug ORG	0.9877	0.9896	0.9982	**0.9998**	0.9997	0.9999
Aug GAN	**0.9902**	**0.9984**	**0.9996**	0.9990	**0.9998**	**0.9999**

to *Aug ORG* where flipping algorithm is independent of any training. As the number of images, extrapolated images were increasing until 750 where the improvement reached a plateau as the classifier became less hungry to data. *Aug ORG* and *GAN* performed approximately equally in the region between the two extremes (100, 1300) with the difference becoming more obvious as k increases (see 750, 1000, and 1300). Finally, the red line represents *Aug GAN* where the random flipping is applied online on the combined real and synthetically-generated images ending up with interpolation and extrapolation happening simultaneously. As can be seen in the figure, this mode outperformed all other modes. The amount of improvement was the largest at 250 and 500 as the classifier was in need for positive data, and became smaller as the classifier has seen enough samples at 750 and 1000). The last point at 1300 was less performing than *Aug ORG* by a negligible amount. The Area Under the ROC Curve (AUC) was used as an additional metric to compare the performance with the different modes and training sizes. AUC values are reported in Table 3, the highlighted values are for the highest of the corresponding size. It can be easily seen from the table that *Aug GAN* outperforms other modes where the highest improvement was for size 250 with 0.0136 over *ORG* mode while the best improvement for *Aug ORG* was for the same size by 0.0047 over *ORG* mode. Moreover, to analyse the distribution of the synthetically-generated images and to compare it to the distribution of the real images, t-SNE was used to reduce the dimensionality of the image space by moving to the 2D feature space. Figure 10 can be used to visualize the distributions for one case at $k = 500$. The algorithm was run for a maximum number of iterations of 4000 and 250 as the perplexity. Similarly, Figure 11 shows the distributions of real and fake masses along with normal tissue patches. It should be noted that this algorithm uses random initialization every time it is run, as a result, it might show different allocations for the samples in the figure for different runs. The input for the algorithm is the patches in the original space (128×128) for both real and synthetic images.

5.3. Experimental Settings

All models were built using Pytorch [3] package by Paszke et al. (2017) with the support of online augmentation. Training a DCGAN then a classifier took on average two hours on NVIDIA TITAN X with 12 GB RAM using CUDA ver.9.0. The generator of the DCGAN had 5M parameters, while the discriminator had 8.9M. All these experiments were carried out at VICOROB lab at the University of Girona using a workstation running Linux Ubuntu 18.04.

6. Discussion

Regarding the results for lesion simulation, it has been shown that DCGAN could generate images that have considerable realism and diversity by training the DCGAN on a dataset that has a sufficient number of examples. The generator could capture the distribution of the real images (p_x) and generate samples that are sampled from p_{gen} which is close to the original one. Figure 12 in Annex 9.1 can be used for a qualitative evaluation of the generated images. This figure shows one real batch of 64 mass lesions (top) along with the same number of synthetic ones (bottom). The size of the training set was 4536 mass and micro calcification lesions. For training details, see section 4.3. By comparing the top and the bottom batches, it can be seen that the generator has learned how to generate mass patches as well as mass accompanied with calcification (see real lesion (3,1) and fake (6,7)). Furthermore, this figure shows that the synthetic batch has reasonable diversity ending up in lesions with different shapes and contrast levels. Some of the shown synthetic examples seem to contain either mass only (see (1,5), (6,4), (7,3)), calcification only (see (6,2) and (8,2)), or a combination of mass and calcification (see (5,7), (4,2)). While other works showed how much observers were fooled when distinguishing real among fake images, in this work we used FID in Figure 8 as an objective evaluation method where the generated images had similar distributions of feature maps at the 2048-unit layer of Inception-v3. Some oscillations in FID values appear due to G being learning. As was mentioned in the talk of Goodfellow (2016), using

[3] code and trained generators are available at `https://github.com/Basel1991/Projects/tree/master/master_thesis`

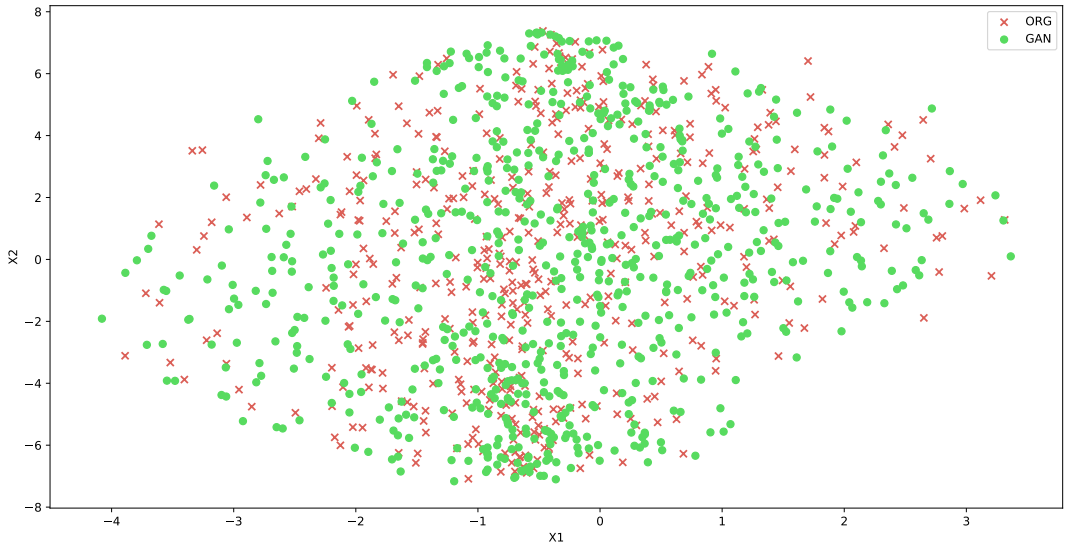


Figure 10: t-SNE analysis for real (red x) and generated (green circles) mass patches distributions, where X1 and X2 axes represent the first and second t-SNE components, respectively.

different kernels between the discriminator and generator along with long training was useful to remove the checkerboard effect and improve the diversity. With respect to evaluating the simulated lesions, we proposed a framework where we train the DCGAN on different-size subsets of the mass dataset (inspired by the work of Frid-Adar et al. (2018)), the trained generators were used to generate synthetic lesions that were used to augment an imbalanced classification problem (normal tissue as the dominant negative class). It was shown that the improvement DCGAN introduced was related to the dataset size. The synthetic images did not improve the performance at a very early stage (very small dataset of 100 images), it did not cause any harm at this stage though. However, the improvement increased with the size of the training set. Moreover, and in line with the results of Bowles et al. (2018), the generated images did some harm for the classifier performance where the F1 score dropped when using the largest subset (P_{1300}) images for training, this suggests that there is a limit for the training set size to assure non-harmful GANs. Aligned with what was interestingly mentioned in Bowles et al. (2018), applying traditional flipping (online and random) method on the real and synthetic images (positive and negative classes) was powerful enough to make DCGAN-generated images helpful regardless the size of the training set (see the red line in Figure 9 for after augmentation and the green one for without augmentation). Additionally, we could show that using the real images only, increasing the size of the training set had a similar impact of enhancing the F1 score but with a much smaller rate with a tipping point where the improvement stops. The distribution of the generated images was analysed and compared to real ones in Figure 10 where it is clear that synthetic images support the distribution of the real ones by filling the gaps in a realistic way as opposed to naive methods which do the averaging of features as in SMOTE and its variations. Figure 11 shows the distribution of real masses, synthetic masses, and normal tissue patches. This figure can show that by using synthetic images, the classifier can generalise more by seeing more examples sampled from the distribution of the minority class (a linear boundary can separate the two distributions). On the one hand, DCGAN could detect the features of the main distribution giving less support to outliers (see the arrow in Figure 11), traditional flipping, on the other hand, does not have the ability to distinguish between inliers (main distribution) and outliers (see the green distribution around the dotted arrow in Figure 13 in Annex 9.3) which can be linked to the improvement in *Aug GAN* over all other methods in Figure 9 and Table 3. Matches, where at least one real and one synthetic samples align perfectly in the t-SNE space (see the solid arrow in Figure 13), are more common in traditional augmentation than in synthetic images due to the fact that the generator does not see the training images.

7. Conclusions

In this study, we used a modified version of DCGAN to generate realistic mammographic lesions with dimensions 128 × 128 pixels that have acceptable diversity. To see the effect of using these synthetically-generated images in action, we simulated an environment where a dataset of mass lesions (as the positive class) and normal tissue (as the negative class) had to be classified with an imbalance ratio of 10. The classification performance was evaluated using F1 score and AUC at six different sizes of the positive dataset

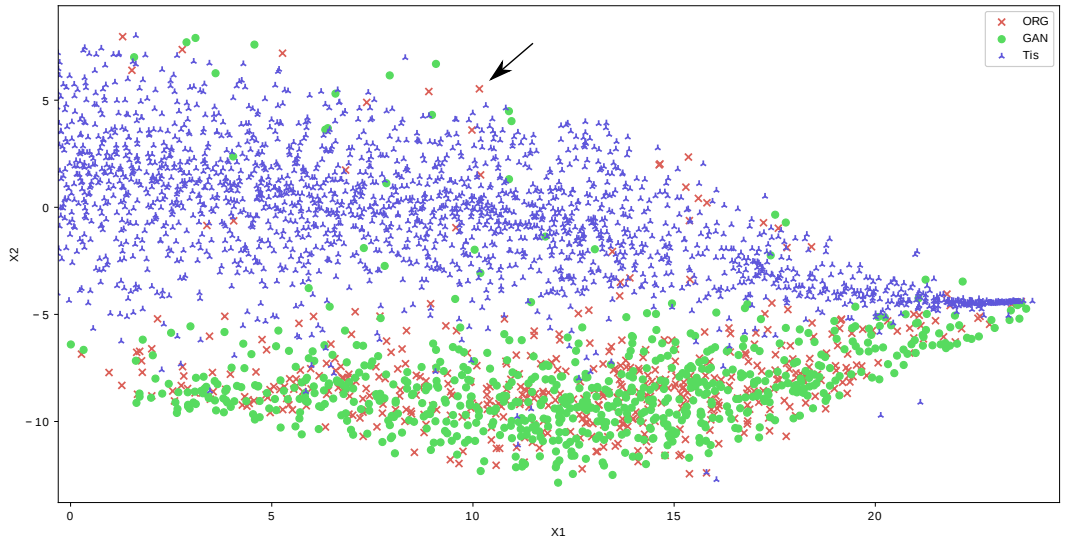


Figure 11: t-SNE analysis for real (red cross) and generated (green circle) mass as well as normal tissue (purple triangle, the negative majority class) patches distributions, where X1 and X2 axes represent the first and second t-SNE components, respectively.

(100, 250, 500, 750, 1000, 1300) keeping the same imbalance ratio, and validated using 3-fold cross validation. We could show that GANs-generated images when used along with online random horizontal then vertical reflection (named as *Aug GAN*) were never harmful and could provide a significant improvement which is higher than when using GAN or flipping individually. This improvement was by the fact that at each size of the training set, *AUG GAN* mode was higher than all other modes resulting in an F1 improvement of approximately (2%, 9%, 8%, 2%, 2%, 2%) over using real images only and approximately (0%, 6%, 2%, 0.5%, 0%, 0%) over using flipping only. Regarding AUC, we could achieve a max improvement of 0.013 over using real images only. Moreover, using synthetic images only as augmentation, there was a limitation at the very small or very large size of the training set where there was either no improvement or a drop in the performance, respectively, compared to real images only. Traditional image flipping augmentation did not suffer from such flaws even without the need for training but could not reach the same level of improvement that *Aug GAN* offered. To sum up, GANs are a powerful tool that can be used to generate synthetic images to be used in a variety of applications including augmenting unbalanced classification problems and unlocking realistic unseen images. However, they have to be trained carefully and better be accompanied with traditional flipping augmentation. In the future, we plan to extend our work to see the effect of using the trained generators on supporting mass detection problems using a different dataset (INbreast). Generating larger patches or even complete mammograms can be explored as well. Furthermore, we are collaborating with radiologists from the Autonomous University of Barcelona to get realism evaluations of the generated mass patches.

8. Acknowledgements

Our great gratitude goes to Nvidia for supporting this work by a Titan X GPU. I would like also to thank my supervisors for offering decent infrastructure and dataset as well as directing me to the target. I am in huge debt to Vicorob research institute, especially Mostafa Abubakr Salem (PhD) for his fruitful discussions and valuable suggestions regarding DCGAN training and testing. Special thanks go to Lavsen Dahal (MAIA) and Albert Garcia (PhD) for their continuous support and enlightening insights.

References

Barratt, S., Sharma, R., 2018. A Note on the Inception Score `arXiv:1801.01973`.

Bazzocchi, M., Mazzarella, F., Del Frate, C., Girometti, F., Zuiani, C., 2007. CAD systems for mammography: a real opportunity? A review of the literature. La radiologia medica 112, 329–353.

Bowles, C., Chen, L., Guerrero, R., Bentley, P., Gunn, R., Hammers, A., Dickie, D.A., Valdés Hernández, M., Wardlaw, J., Rueckert, D., 2018. GAN Augmentation: Augmenting Training Data using Generative Adversarial Networks `arXiv:1810.10863`.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A., 2018. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians 68, 394–424. doi:`10.3322/caac.21492`.

Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W., 2002. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, 321–357. URL: `https://www.scopus.com/inward/record.uri?eid=2-s2.0-0346586663&partnerID=40&md5=dfb419b8460388447758f9c7f8c2a103`. cited By 4652.

Douzas, G., Bacao, F., 2018. Effective data generation for imbalanced learning using conditional generative adversarial networks. Expert Systems with Applications 91,

464 – 471. URL: http://www.sciencedirect.com/science/article/pii/S0957417417306346, doi:10.1016/j.eswa.2017.09.030.

Elmore, J.G., Wells, C.K., Lee, C.H., Howard, D.H., Feinstein, A.R., 1994. Variability in radiologists' interpretations of mammograms. New England Journal of Medicine 331, 1493–1499. doi:10.1056/NEJM199412013312206. pMID: 7969300.

Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., Greenspan, H., 2018. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. Neurocomputing 321, 321 – 331. URL: http://www.sciencedirect.com/science/article/pii/S0925231218310749, doi:10.1016/j.neucom.2018.09.013.

Goodfellow, I., 2016. NIPS 2016 Tutorial: Generative Adversarial Networks arXiv:1701.00160.

Halling-Brown, M.D., Patel, P.T.L.M.N., Warren, L.M., Mackenzie, A., Young, K.C., 2014. The oncology medical image database (omi-db), in: Proc. SPIE 9039 Medical Imaging 2014: PACS and Imaging Informatics: Next Generation and Innovations. doi:10.1117/12.2041674.

He, H., Garcia, E.A., 2009. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering 21, 1263–1284. doi:10.1109/TKDE.2008.239.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, in: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30. Curran Associates, Inc., pp. 6626–6637. http://papers.nips.cc/paper/7240-gans-trained-by-a-two-time-scale-update-rule.

J. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengioa, Y., 2014. Generative Adversarial Nets, in: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 27. Curran Associates, Inc., pp. 2672–2680. URL: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Medical Image Analysis 36, 61 – 78. URL: http://www.sciencedirect.com/science/article/pii/S1361841516301839, doi:10.1016/j.media.2016.10.004.

Karras, T., Aila, T., Laine, S., Lehtinen, J., 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation, in: International Conference on Learning Representations. URL: https://openreview.net/forum?id=Hk99zCeAb.

Kingma, D.P., Ba, J., 2014. Adam: A Method for Stochastic Optimization arXiv:arXiv:1412.6980.

Kohli, A., Jha, S., 2018. Why CAD Failed in Mammography. Journal of the American College of Radiology 15, 535 – 537. URL: http://www.sciencedirect.com/science/article/pii/S1546144017316745, doi:10.1016/j.jacr.2017.12.029. Data Science: Big Data Machine Learning and Artificial Intelligence.

Korkinof, D., Rijken, T., O'Neill, M., Yearsley, J., Harvey, H., Glocker, B., 2018. High-Resolution Mammogram Synthesis using Progressive Generative Adversarial Networks arXiv:1807.03401v1.

Le, E., Wang, Y., Huang, Y., c, S.H., Gilbert, F., 2019. Artificial Intelligence in Breast Imaing. Clinical Radiology 74, 357–366. doi:10.1016/j.crad.2019.02.006.

Lei Ba, J., Kiros, J.R., Hinton, G.E., 2016. Layer Normalization arXiv:1607.06450.

Lucic, M., Kurach, K., Michalski, M., Gelly, S., Bousquet, O., 2018. Are GANs Created Equal? A Large-Scale Study, in: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 31. Curran Associates, Inc., pp. 700–709. URL: http://papers.nips.cc/paper/7350-are-gans-created-equal-a-large-scale-study.pdf.

van der Maaten, L., Hinton, G., 2008. Visualizing Data using t-SNE. Journal of Machine Learning Research 9, 2579–2605. URL: http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf.

Malvezzi, M., Carioli, G., Bertuccio, P., Boffetta, P., Levi, F., La Vecchia, C., Negri, E., 2019. European cancer mortality predictions for the year 2019 with focus on breast cancer. Annals of Oncology 0, 1–7. doi:10.1093/annonc/mdz051.

Mirza, M., Osindero, S., 2014. Conditional Generative Adversarial Nets arXiv:arXiv:1411.1784.

Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y., 2018. Spectral normalization for generative adversarial networks, in: International Conference on Learning Representations. URL: https://openreview.net/forum?id=B1QRgziT-.

Moreira, I.C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M.J., Cardoso, J.S., 2012. INbreast: Toward a Full-field Digital Mammographic Database. Academic Radiology 19, 236 – 248. URL: http://www.sciencedirect.com/science/article/pii/S107663321100451X, doi:10.1016/j.acra.2011.09.014.

Orel, S.G., Kay, N., Reynolds, C., Sullivan, D.C., 1999. Birads categorization as a predictor of malignancy. Radiology 211, 845–850. doi:10.1148/radiology.211.3.r99jn31845. pMID: 10352614.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in PyTorch, in: NIPS Autodiff Workshop.

Radford, A., Metz, L., Chintala, S., 2016. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks, in: 2016 International Conference on Learning Representations (ICLR).

Salehinejad, H., Valaee, S., Dowdell, T., Colak, E., Barfett, J., 2018. Generalization of Deep Neural Networks for Chest Pathology Classification in X-Rays Using Generative Adversarial Networks, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 990–994. doi:10.1109/ICASSP.2018.8461430.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved Techniques for Training GANs arXiv:1606.03498.

Samardar, P., de Paredes, E.S., Grimes, M.M., Wilson, J.D., 2002. Focal Asymmetric Densities Seen at Mammography: US and Pathologic Correlation. RadioGraphics 22, 19–33. doi:10.1148/radiographics.22.1.g02ja2219. pMID: 11796895.

Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z., 2016. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the Inception Architecture for Computer Vision, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Wu, E., Wu, K., Cox, D., Lotter, W., 2018. Conditional Infilling GANs for Data Augmentation in Mammogram Classification, in: Stoyanov, D., Taylor, Z., Kainz, B., Maicas, G., Beichel, R.R., Martel, A., Maier-Hein, L., Bhatia, K., Vercauteren, T., Oktay, O., Carneiro, G., Bradley, A.P., Nascimento, J., Min, H., Brown, M.S., Jacobs, C., Lassen-Schmidt, B., Mori, K., Petersen, J., San José Estépar, R., Schmidt-Richberg, A., Veiga, C. (Eds.), Image Analysis for Moving Organ, Breast, and Thoracic Images, Springer International Publishing, Cham. pp. 98–106. doi:10.1007/978-3-030-00946-5_11.

9. Annex

In these annexes, we show the outcome of four experiments. First, we present a real and a fake batches of 64 images each, where the fake ones were generated via a DCGAN trained on the complete training dataset (4536 mass + calcification). Second, a figure that shows the progress of training the DCGAN accompanied with the loss plot and a batch of 4 fake images enhancement during training. Third, we include t-SNE analysis for *Aug ORG* showing real and augmented masses as well as normal tissue in the 2D feature space of t-SNE. Fourth and last, we show 25 random samples generated from six different generators trained on $(100, 250, 500, 750, 1000, 1300)$ images individually and we compare the quality and diversity.

9.1. A Real And A Fake Batch

In this section, using Figure 12, we show one real batch of 64 mass lesions (top) along with the same number of fake ones (bottom) generated by a generator that was trained on 4536 images (mass + calcification).

9.2. GAN Training Progress

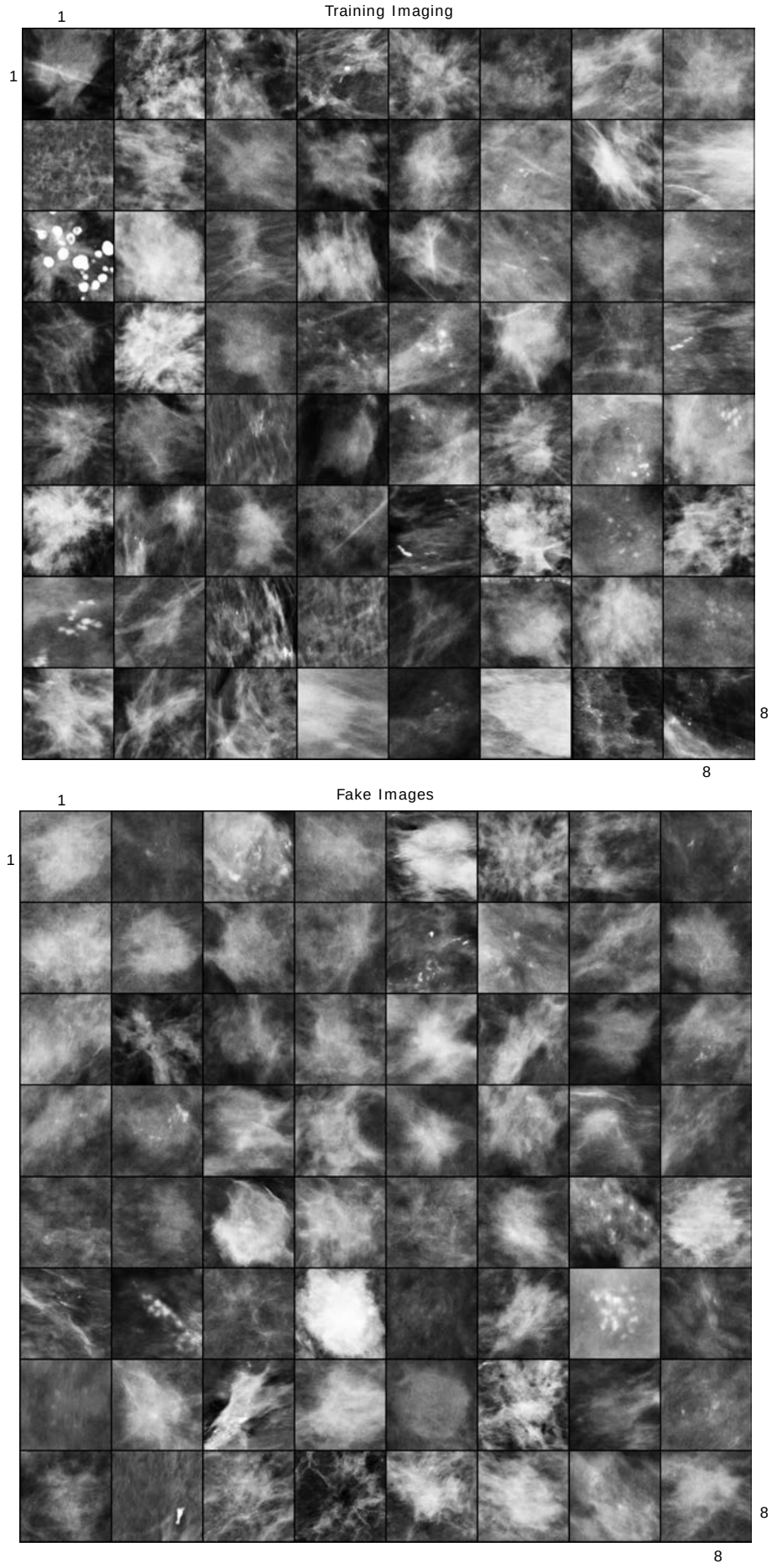
Here we show the progress of G and D loss during training using the mass + calcification dataset of size 4536. The reflection on images realism and diversity is explored in Figure 14 where it shows G and D average loss along with samples of 4 images generated from a fixed noise batch. By looking at the beginning of the plot (iteration 0), the loss of G starts high because it starts with random output which is relatively easy even for an inexperienced discriminator to realise that it is not real. In this case, the output of D for the fake input is very low (low realism probability). At iteration 0, D has just started to learn, however, the process of distinguishing real patches among real ones is considered easy, however, this becomes tougher when G starts learning. During iterations 0 to 20,000, G is learning from its mistakes by modifying the weights relatively to D output and competing with D which has a merely-constant average loss. A large drop of more than 70% in FID is due to the large gradients of NS loss (see Figure 5). it can be seen from the difference in quality between the batches at epoch 140 and epoch 420 where the checkerboard effect was removed with an increase in realism and a decrease in FID. At iteration 25,000, D becomes almost professional and no more gets fooled by G output (D loss is monotonically decreasing), on the contrary, G loss starts increasing but keeps improving (see the image at epoch 700 where the lesions have been improved in terms of contrast and size). Iterations from 50K until 70K have lesser impact due to the learning rate here being too small compared to the early stages, still, this period had a subtle contribution to improvements in image diversity. By looking at FID values, it is clear that the decrease was exponentially decaying (along with the learning rate). At the end, the generator seems as it has lost the game by getting the loss settles at a relatively high value. It should be kept in mind here that the real label was randomly chosen every epoch which had the impact of the oscillations in the losses. The discriminator could keep detecting real images among fake ones ending up winning the game, this is fine as G had enough time to learn.

9.3. t-SNE Analysis for Aug ORG

Figure 13 shows the embeddings for 500 real mass patches, 750 with random flipping, and 5K normal tissue (negative majority).

9.4. A Sample From Each G_k

I this section, the aim is to evaluate subjectively the effect of increasing the training set size on the realism and diversity of the generated images. Figure 15 shows six batches containing 25 images each, batches from top to bottom and left to right were generated by $G_{100}, G_{250}, G_{500}, G_{750}, G_{1000}, G_{1300}$ (see section 4.6). Starting with 100, this batch shows a low diversity (low recall) in lesions shapes with some similarity between lesions (see batch 100, (1,4) and (3,1), (3,3) and (4,3)). This suggests a mode collapse situation with the realism being not high. Moving on to batch 250, it is noticeable that lesions here have more contrast than before with some new shapes, however, there is still some patterns that are repeated between lesions (see batch 250 (1,1), (3,3) and (3,4)). Diversity keeps improving as well as realism when reaching to 500 and 750 where it is hard to detect such patterns. It can be seen that at 750 the generator has learned to sample lesions with more detailed architectures than before in batch 100. Batches 1000 and 1300 are where the generator starts to generate images that are hard to distinguish from real ones.



1.18

Figure 12: Two batches of real and fake images. The top one is the real batch while the bottom one is the fake one. Indices used here are of the shape (i,j), where i is the row index, j is the column index and the top left being (1,1).

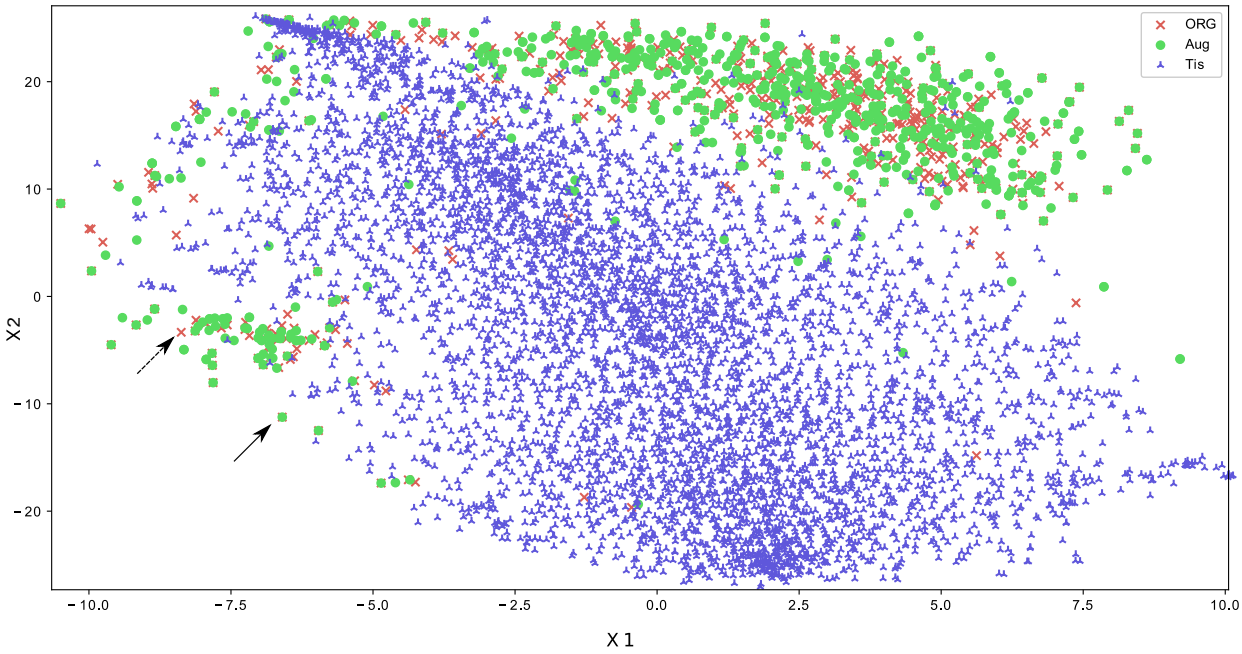


Figure 13: t-SNE distributions for real masses, flipped masses, and normal tissue patches. A red x represents a real mass patch, a green circle represents a flipped mass (horizontal, vertical, both, or none), purple triangles represent normal tissue patch. The dotted arrow points at an outlier while the solid one points at a match.

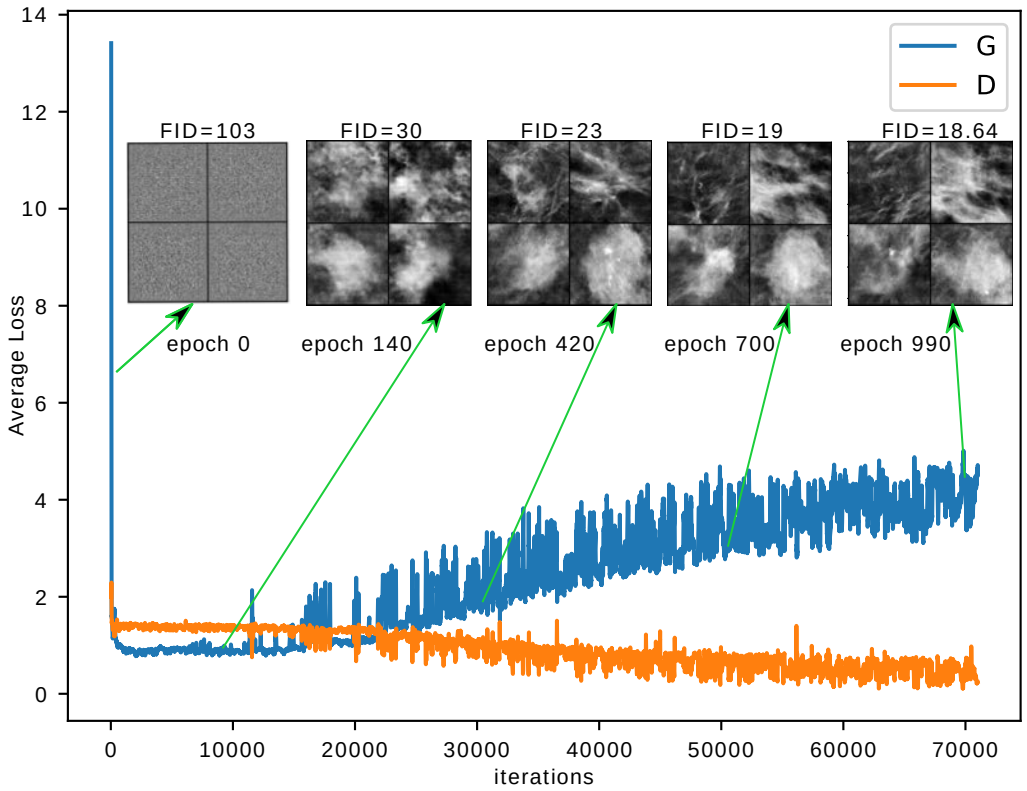


Figure 14: GAN progress, showing 5 batches of generated images at different points in the training process, these images were taken from epochs 0, 140, 420, 700, and 990, where the input was fixed to four latent vectors. The horizontal axis is the training iterations, the vertical one is for DCGAN adversarial loss, see equations (1,2). FID values are approximated and provided for each case.

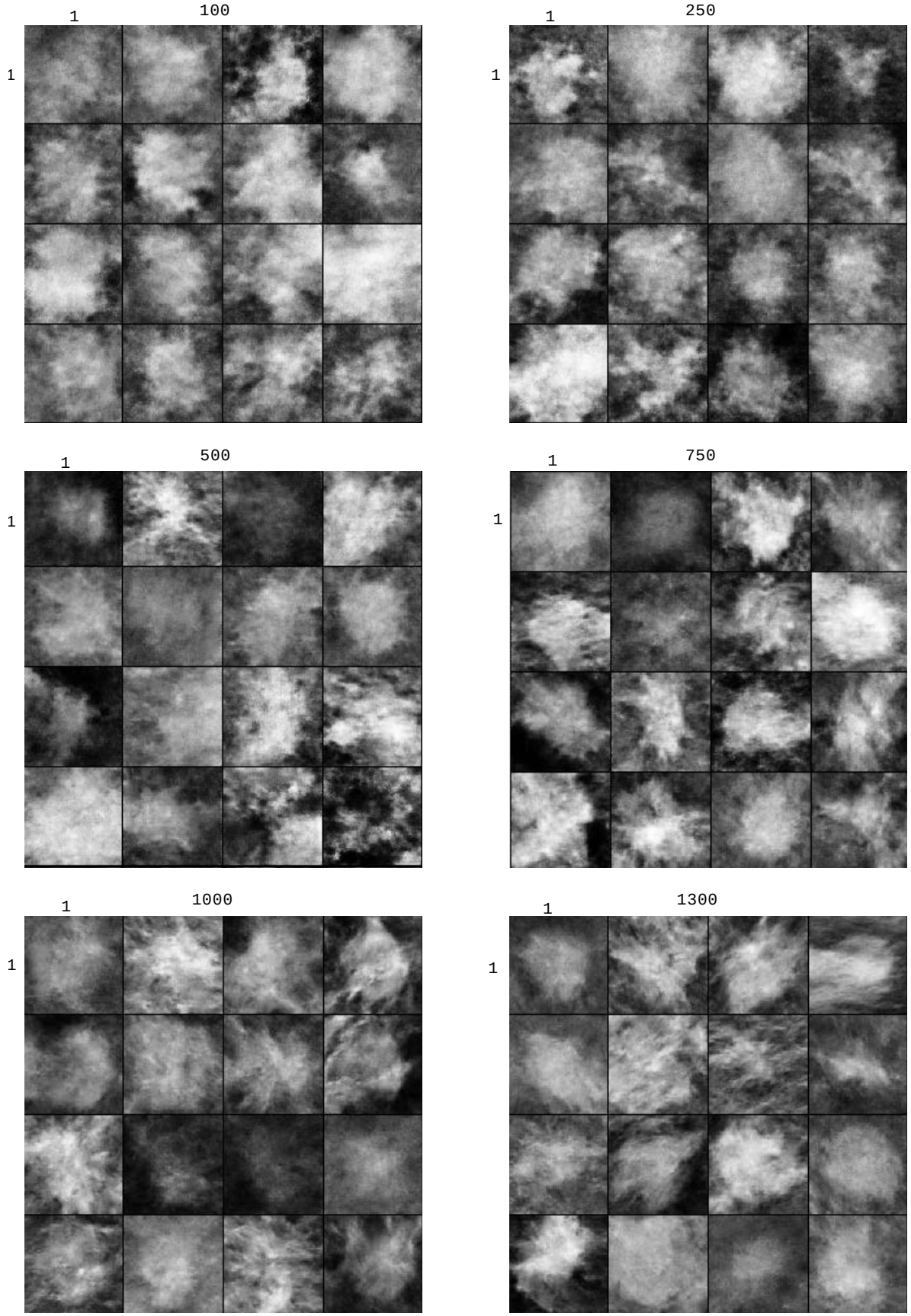
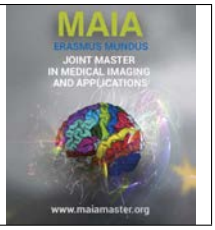


Figure 15: From top to bottom and left to right, 25 random samples generated from G_k : $k = \{100, 250, 500, 750, 1000, 1300\}$. This figure is to show the relationship between image quality and diversity, and number of training images for the DCGAN. Indices used here are of the shape (i,j), where i is the row index, j is the column index and the top left being (1,1).

Medical Imaging and Applications

Master Thesis, June 2019



Deep Learning for Fast Temporal Mammographic Image Registration

Ali Berrada, Robert Marti, Oliver Diaz

Computer Vision and Robotics Lab, University of Girona, Spain

Abstract

The comparison of temporal mammograms helps radiologists in breast cancer diagnosis by analyzing interval changes. The breast's appearance, however, may vary considerably on mammograms taken at different sessions due to several extrinsic and intrinsic factors experienced by the breast. Consequently, a direct comparison of these images may become a complex and exhausting endeavor for the radiologist if no preliminary alignment is done. Therefore, several techniques have been proposed for mammographic image registration. While some degree of success has been documented, these methods tend to be time consuming requiring from several minutes to hours.

With the revolution of deep learning in the field of computer vision and image processing in the past few years, the research community started to leverage the power of neural networks for the medical image registration task. In line with this trend, a recent paper promoted VoxelMorph as a novel, fast and self-learning framework based on convolutional neural networks capable of achieving state-of-the-art registration accuracy while requiring orders-of-magnitude lesser time than traditional techniques.

In this master thesis, we presented VoxelMorphMammography (VMM), an extension of the VoxelMorph framework adapted for the image registration task of temporal mammograms. The conducted experiments involved using two different similarity metrics for the loss function and different regularization weights to constrain the deformation. The evaluation of the registration results were based on a variety of quantitative and qualitative measures: mutual information, subtraction images, warped images and deformation fields. With all these indicators jointly taken into consideration, we found that VMM is able to achieve results comparable to a state-of-the-art method based on B-splines. Furthermore, VMM registered each test image pair in less than 100ms. Finally, we call for more research and experimentation to be done to provide quality assurance in using deep learning approaches for mammographic image registration.

Keywords: deformable image registration, mammography, deep learning

1. Introduction

Breast cancer is the most common form of tumor that affects women worldwide (Bray et al., 2018). In 2015, the world age-standardized rate of breast cancer incidence in Spain was 65.2 per 100,000 women (Galceran et al., 2017). It is also the leading cause of cancer-related death among women (Fitzmaurice et al., 2017). However, with increased awareness about the importance of early diagnosis, breast cancer can be effectively treated before turning into a life-ending disease if detected in an early stage. For this reason, breast cancer screening programs are performed in most developed countries (Schopper and de Wolf, 2009).

Different imaging modalities can be used for the screening. Yet, due to its affordability, accessibility and high sensitivity, mammography is the most undertaken technique for breast imaging (Sree et al., 2011). Particularly, screening mammography, as a preventive measure for early detection of cancerous lesions and carcinoma in situ, enables the identification of non-palpable lesions which will be otherwise undetected during a physical breast examination (Ekeh et al., 2000).

Mammography utilizes low-energy X-rays to build an image of the breast called a mammogram. The mammogram can be obtained from different angles. In a routine screening exam, two standard views of the breast are taken: one from above called the cranial-

caudal (CC) view and one from an oblique projection called the mediolateral-oblique (MLO) view (Wei et al., 2016). During these acquisitions, the breast remains compressed to minimize subject's motion, dose and scattered radiation. Breast tumors have attenuation properties similar to dense tissue, thus making their detection a challenge.

Overtime, the screening population has a record of two or more mammograms either as part of a regular and preventive screening plan or through a periodic mammography surveillance for probably benign lesions. The availability of a pair of breast images acquired at different times helps the radiologist to look for any abnormal changes in breast tissues. This is a highly productive strategy as it has been found to yield to improved cancer detection (Bassett et al., 1994). Additionally, being able to switch between a pair of mammograms on the same monitor optimizes the radiologist's perception of lesion growth (van Engeland et al., 2003). However, the deformable characteristic of the breast makes the comparison of temporal mammograms a challenging task and often exhausting when the radiologist has to analyze multiple images. The mammography acquisition parameters, the positioning and compression of the breast, and anatomical changes are among the main factors that affect how the breast appears in a mammogram (Richard et al., 2006). As a consequence, many research works have investigated strategies to find a reliable non-rigid transformation technique to align time-sequenced mammograms (Abdel-Nasser et al., 2016; Marias et al., 2005, 1999; Timp et al., 2005). This alignment problem is known as image registration.

Image registration is the process of aligning one image to another image. This is achieved by finding a transformation function that establishes the pixel correspondences between two images. In common terminology, the image to be registered is called the moving (or source) image, while the image to be aligned to is called the fixed (or target) image. A warped image is the image obtained by warping the source image using the transformation function.

Two broad categories of image registration can be defined based on the transformation model applied: rigid registration and non-rigid, or deformable, registration. Rigid registration employs global linear transformations such the Euclidean and affine transforms and therefore cannot model local geometrical distortions between the images. On the other hand, non-rigid registration enables a more complex and non-uniform mapping of pixels when warping the moving image to the fixed image. The B-Spline transform is an example of a popular model that caters for local image deformations. Often, a deformable registration will be preceded by a relatively fast affine registration for global alignment and better initialization.

Image registration requires a metric to measure similarity that guides the registration process. Commonly used metrics include Mean Squares Difference (MSD), Normalized Cross Correlation (NCC) and Mutual Information (MI). The choice of one metric depends on the characteristics of images. For example, MSD can be used when images have the same range of intensity values while MI is used to measure the amount of information that the registered image contains about the target image.

In the medical practice, several diagnostic and therapeutic applications, including mammography, require deformable registration as organ structures change shape or position between different scans. Consequently, the last decade has seen a surge of publications in this domain. While many of the traditional and common methods achieve high registration accuracy, they tend to suffer from time complexity requiring from several minutes to hours for a single registration (Klein et al., 2009). This is a result of these methods solving an optimization problem for every pair of images computing the non-linear correspondences. Algorithms adapted for graphics processing unit (GPU) can considerably reduce the registration time but eventually every registration depends on GPU (Modat et al., 2010).

With the revolution of deep learning in the field of computer vision and image processing in recent years, the research community started to leverage the power of neural networks for the medical image registration task. Several papers presented a supervised learning approach, hence requiring ground truth data (Cao et al., 2017; Krebs et al., 2017; Sokooti et al., 2017; Yang et al., 2017). In these methods, the ground truths correspond to warp fields obtained either by registering images using traditional methods or by deforming images using devised transformations. The former clearly presents inconveniences and limits the diversity of deformations that can be learned whereas the latter cannot simulate certain deformations that happen in reality. Only few work has been done to promote a self-learning framework but come with experimental limitations (Li and Fan, 2017; de Vos et al., 2017). A very recent paper proposed VoxelMorph, a fast and unsupervised learning framework for deformable medical image registration (Balakrishnan et al., 2018). The authors claimed that their algorithm, applied to 3D magnetic resonance (MR) brain images, results in registration accuracy similar to the state-of-the-art methods while execution taking orders-of-magnitude less time: less than a minute on a central processing unit (CPU) and under a second on a GPU.

From our survey of available literature, we found multiple papers dealing with the alignment problem of mammograms; yet, none of them is applying machine learning techniques. Deep learning is increasingly used to detect and classify breast masses or calcifications in mammographic images but not to register them. This present work, therefore, aims at filling this gap by

studying VoxelMorph and assessing the possibility of this framework in replacing known non-learning-based methods.

We named our adaptation of VoxelMorph for mammogram images as VoxelMorphMammography (VMM) to give credit to the original work by Balakrishnan et al. (2018).

2. State of the art

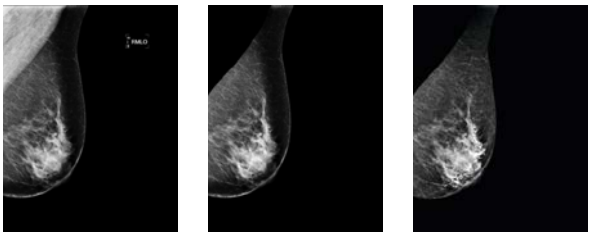
Breast image registration is a an ongoing challenging task that has been researched for years. The need for registration is not limited to temporal mammography, but extends to bilateral mammography, breast MR imaging (MRI), and even mutli-modal such as the fusion of breast MRI and mammography.

The literature abounds with several papers proposing different methods for tackling the challenge of breast image registration and harvest its clinical benefits (Boehler et al., 2012; Hipwell et al., 2016).

Of the simplest yet effective ways to reduce misalignment between images is to apply an affine transformation. While it cannot correct for local deformations, it can provide significant initial alignment. In their study, Pinto Pereira et al. (2010) found that affine registration was almost as accurate as trained experts in matching landmarks on mammograms. On the other hand, Mertzanidou et al. (2012) advanced a volume-preserving multi-modal registration algorithm based on the affine transform to help radiologists in breast cancer diagnosis. Nevertheless, the affine transformation is often a precursor for more advanced image registration algorithms. For instance, Rueckert et al. (1999) proposed their method to register contrast-enhanced breast MR images consisting of a combination of an affine transformation and a free-form deformation (FFD) based on B-splines. The first finds the global motion of the breast while the latter models local shifts. The results were superior when compared to using rigid transformations only. Rohlfing et al. (2003) extended on Rueckert et al. (1999)'s work by adding a regularization term which helped in volume preservation of breast tissues.

Other methods seek to extract features from mammograms to model the deformations. For instance, Marias et al. (1999)'s method consists of identifying specific landmarks along the breast boundary which will be the basis of matching mammograms using an interpolation function based on thin-plate splines. On the other hand, Wai and Brady (2005) achieved breast alignment in mammograms by constructing and matching anatomy-mimicking curvilinear coordinates.

More techniques can be found in the literature revolving around similar concepts. However, we could not find any paper presenting a machine learning approach for aligning breast images. Neural networks have the potential to considerably reduce the image registration time which tends to be long with traditional methods (Klein et al., 2009).



(a) Original mammogram (b) Pectoral muscle removed (c) Peripheral enhancement

Figure 1: Pre-processing steps.

3. Material and methods

3.1. Dataset

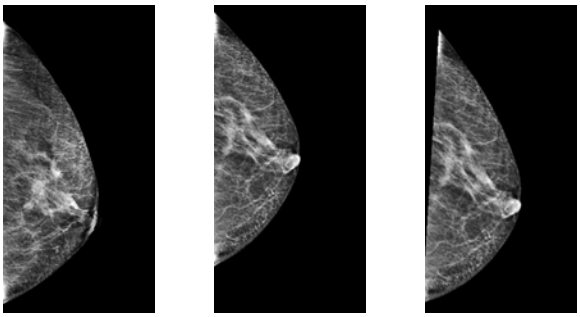
The dataset used is a subset of a larger database of full-field digital mammograms made accessible to the members of the Computer Vision and Robotics Lab at the University of Girona, Spain. The acquisition was done using a Hologic Selenia system at the resolution of 70 micron per pixel. The images are 12-bit depth and their spatial resolution is either 4096x3328 pixels or 2560x3328 pixels. For our experiments, we have arranged 160 pairs of mammograms of which 133 pairs are of the CC view while the rest (i.e. 27) are of the MLO view. The CC-view mammograms will be used for training and testing to understand the effect of the framework parameters on the registration results. On the other hand, the MLO-view mammograms will be used for testing networks trained solely on CC-view images to evaluate the applicability of such scenario. In every pair, the images are of the same breast with 1 or 2 years difference between the acquisitions.

3.1.1. Pre-processing

The pre-processing of the images follows the pipeline done by Tortajada et al. (2014). The goal is to reduce the variability between the sequence of mammograms for easing the registration step. Practically, the breast was segmented using simple thresholding while the pectoral muscle was suppressed using an automatic algorithm (Kwok et al., 2004). Additionally, a peripheral enhancement technique was employed to compensate for thickness inconsistencies at the breast periphery as devised by Tortajada et al. (2012). Figure 1 shows the result of each pre-processing step applied to a sample mammogram.

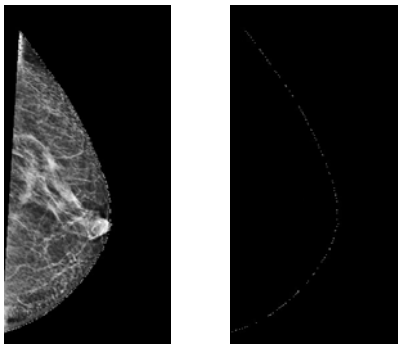
The images were downsampled to 512x448 pixels to reduce their memory size and speed up the training phase.

Finally, we performed an affine registration for each mammogram pair using elastix to obtain a global alignment (see Figure 2). The main configuration of this



(a) Fixed image (b) Moving image (c) Warped image

Figure 2: Affine registration: (c) is the result of warping (b) using our affine transformation model.



(a) Warped image (b) Breast region removed from (a)

Figure 3: Unwanted white pixels along breast boundary generated when selecting a B-spline of order greater than 1 for the final resampling interpolation.

registration was a 6-level multiresolution scheme, mutual information for the cost function, adaptive stochastic gradient descent for the optimizer, and first order B-spline for the interpolator. At the final step of the registration when the pixel correspondences between two images are found, the moving image is transformed via resampling and this requires interpolation. We used a B-spline of order 1 as the resampling interpolator since higher orders generated unwanted white pixels along the breast boundary as can be seen in Figure 3.

3.2. Environmental Setup

The data processing, code development and elastix-driven registrations were carried on a personal computer powered with an Intel Core i5-3230M processor cloaked at 2.60 GHz and equipped with 12 GB of RAM.

For training the neural network we have used Colaboratory, a free cloud service for machine learning (Google, 2019). It is a setup-free Jupyter notebook environment that can be accessed from the browser. The computations are performed on a virtual machine operated by Ubuntu version 18.04.2 LTS and powered with a NVIDIA Tesla K80 and 12 GB of RAM. Colaboratory has been promoted as an effective platform for deep learning acceleration by Carneiro et al. (2018) and was featured in several research projects (Bodhwani et al., 2019; Roy et al., 2019; Satılmış et al., 2018).

For affine and B-spline registrations, we used a popular toolbox for intensity-based medical image registration called elastix (Klein et al., 2010). Launching an image registration task with elastix is quick and simple as it allows the user to define and tune the registration strategy (multiresolution scheme, transformation function, loss function, optimizer, etc.) in a parameter text file.

3.3. Baseline Method

Diez et al. (2011) quantitatively measured the accuracy of different state-of-the-art image registration methods in registering temporal mammograms. The results of the experiments were in favor for the method based on B-splines and adopting a multiresolution registration paradigm. Therefore, we chose this method to be our baseline for comparing with VMM.

Practically, we performed the registration using elastix. The main parameters were: 4-level multiresolution scheme, mutual information for the cost function, adaptive stochastic gradient descent for the optimizer, 20mm for the B-spline's grid spacing, and first order B-spline for the interpolator and resampling interpolator.

3.4. VoxelMorph (VM)

3.4.1. Overview

Balakrishnan et al. (2018) devised an unsupervised learning-based framework for non-rigid medical image registration called VoxelMorph. Their experiments showed that VoxelMorph is able to achieve performances comparable to those from state-of-the-art methods while requiring much lesser time for performing actual registration tasks.

At the heart of the VoxelMorph framework is a convolutional neural network (CNN) that models a function $g_\theta(f, m) = \phi$, where θ represents the learnable network parameters (specifically, the kernels of the convolutional layers), f and m are respectively the fixed and moving image defined over a spatial domain Ω, and ϕ is the registration or deformation field.

The deformation field ϕ is defined in the fixed image space. In the case of registering 2D images, it is stored as a 2-channel image where the first and second channel respectively specify the pixel displacements along the rows (i.e., negative y-direction) and along the columns (i.e., x-direction).

The framework is summarized in Figure 4. Input images f and m are fed to the θ-parameterized convolutional network to compute ϕ. Using a spatial transform function, m is warped with ϕ. The warped image m' and f are evaluated for similarity by a loss function L. The goal of the training phase is to minimize L by finding the optimal parameters θ. This is done using the stochastic gradient descent method.

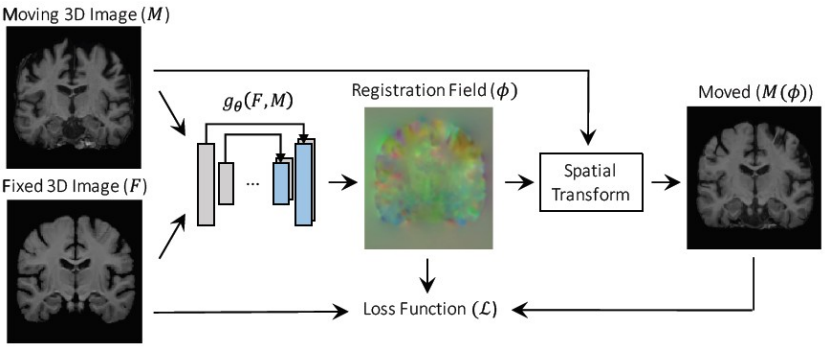


Figure 4: VoxelMorph framework (Balakrishnan et al., 2018)

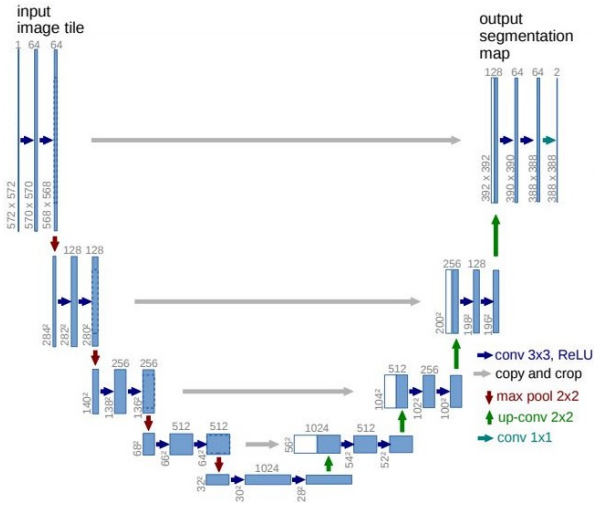


Figure 5: Original U-Net architecture (Ronneberger et al., 2015)

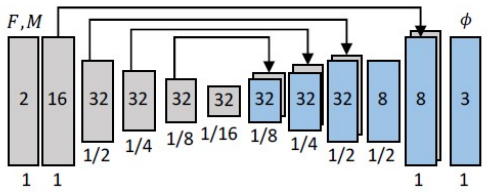


Figure 6: VoxelMorph's CNN proposed at CVPR 2018 (Balakrishnan et al., 2018)

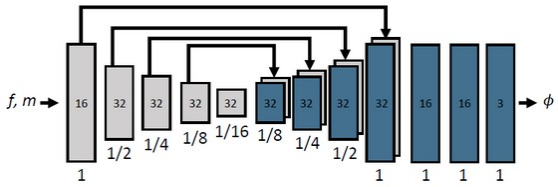


Figure 7: VoxelMorph's CNN proposed at MICCAI 2018 (Balakrishnan et al., 2019)

3.4.2. Network Architecture

VoxelMorph implements a U-Net inspired CNN. U-Net was developed by Ronneberger et al. (2015) for medical image segmentation tasks and has gained a wide popularity within the research community and is often employed for segmentation challenges. The U-Net architecture, as shown in Figure 5, has three main characteristics: a contracting path, an expanding path and skip connections. The contracting path consists of a number of convolutional layers sequentially downsampled by a max pooling layer of stride 2. Through this path, the network learns the important features (context) but not their locations. The resolution is also low at this stage. The expanding path has a reversed effect as it works to progressively propagate context information to higher resolutions until achieving the original resolution. For this, it replaces max pooling operations by transposed convolutions for upsampling. The precise localization of context at each stage of the expanding path is ensured by the skip connections that concatenate the upsampled layers with their corresponding feature maps from the contracting path. All convolutions are activated by a rectified linear unit (ReLU) function.

The VoxelMorph network has certain variations from the original U-Net. First, the input to the network is a single layer made by concatenating m and f. The downsampling, where the resolution is reduced by half at each layer, is the result of using stride 2 convolutions instead of max pooling. The convolutions are followed with a Leaky ReLU layer which mitigates the infamous "Dying ReLU" problem (Maas et al., 2013). Also, the upsampling operations along the expanding path are simply a duplication of rows and columns.

Figure 6 shows the network presented at the 2018 International Conference on Computer Vision and Pattern Recognition (CVPR) while Figure 7 depicts the model used for the 2018 International Conference on Medical Image Computing & Computer Assisted Intervention (MICCAI). The second model is part of the authors' extended work in Balakrishnan et al. (2019) where a diffeomorphic formulation was integrated. The purpose is to generate deformation fields that are smooth and invertible to cater for many scientific applications where it is important to analyze these fields.

3.4.3. Loss Functions

VoxelMorph does not make use of ground truth registration fields during the network training. The unsupervised learning is achieved by evaluating the model through the loss function L (Equation 1). It consists of two components: L_{sim} which estimates dissimilarity between f and m', and L_{smooth} which introduces diffusion regularization to ϕ. The second term helps in producing smooth warp fields which may otherwise reveal unrealistic physical deformations since L_{sim} will push m' to verge on f. The penalization of irregular local spatial variations in ϕ is controlled by a weighing parameter λ.

$$L(f, m, \phi) = L_{sim}(f, m') + \lambda L_{smooth}(\phi) \tag{1}$$

Two intensity-based similarity methods for L_{sim} are available in VoxelMorph:

1. Mutual Squared Error (MSE): This metric quantifies dissimilarity since it computes the intensity difference between corresponding pixels from two images. Therefore, it can be used when f and m have similar intensity distributions, otherwise the registration performance will be suboptimal (Bağcı et al., 2010). Nevertheless, intensity variations can be minimized by normalizing the images. The formula is defined in Equation 2.
2. Local Cross-Correlation (LCC): Cross-correlation computes a coefficient that reveals the degree of similarity between f and m': the higher the cross-correlation value is, the higher the similarity is. Consequently, when using this metric, the additive inverse is taken since the algorithm tries to minimize the error (i.e. the dissimilarity between images); therefore, L_{sim} is formulated as: $L_{sim} = -LCC(f, m')$. Contrary to MSE, cross-correlation is not affected by variations in intensity scale between images (Roshni and Revathy, 2008). In the standard intensity-based cross-correlation formula, the mean intensity of the whole images are subtracted. Instead, VoxelMorph employs a local cross-correlation formulation where the local mean from a pixel's neighborhood is computed (Equation 3).

$$MSE(f, m') = \frac{1}{|\Omega|} \sum_{p \epsilon \Omega} (f(p) - m'(p))^2 \tag{2}$$

$$LCC(f, m') = \sum_{p \epsilon \Omega} \frac{(\sum_{p_i}(f(p_i) - \hat{f}(p))(m'(p_i) - [\hat{m} \circ \phi](p)))^2}{(\sum_{p_i}(f(p_i) - \hat{f}(p))^2)(\sum_{p_i}(m'(p_i) - [\hat{m} \circ \phi](p))^2)} \tag{3}$$

where:

- p_i iterates around p in a window size of 9,
- $\hat{f}(p) = f(p) - \frac{1}{n^2} \sum_{p_i} f(p_i)$,
- $\hat{m}(p) = m(p) - \frac{1}{n^2} \sum_{p_i} m(p_i)$, and
- $\hat{m} \circ \phi$ denotes $\hat{m}$ warped by ϕ.

For L_{smooth}, VoxelMorph employs a diffusion regularizer based on the heat diffusion equation and therefore computes gradients of ϕ (Equation 4). While the gradient of an image reflects the magnitude and direction of changes in intensity, the gradient of ϕ reflects the magnitude and direction of changes in displacement. Abrupt changes therefore indicate irregularities in ϕ whereas gradual changes imply smoothness.

$$L_{smooth}(\phi) = \sum_{p \epsilon \Omega} \|\Delta \phi(p)\|^2 \tag{4}$$

3.4.4. Spatial Transformer

The Loss function evaluates in part the similarity between f and m'. This metric helps the optimization of the network parameters θ. However, the network computes ϕ rather than directly estimating m'. Therefore, after finding ϕ and before computing the loss, m' is obtained by warping m with ϕ. This is achieved by using a transformation function that is inspired by the idea of the spatial transformer networks (Jaderberg et al., 2015). The transformer is a differentiable operation; this allows the loss to be backpropagated during the optimization of the parameterized ϕ.

3.5. VoxelMorphMammography (VMM)

Though experimented on 3D brain MR images, VoxelMorph was designed for generic pairwise medical image registration. In this study, we sought to build upon the original work and provide a framework adapted to registering temporal mammograms and named it VoxelMorphMammography. The changes we brought to the original framework are described in the following subsections.

3.5.1. Network Architecture

We present the VMM neural network in Figure 8. The input to the network is a 2-channel concatenation layer made of f and m. When the convolution operation is to preserve the spatial resolution, a stride of 1 is used, otherwise a stride 2 is used resulting in outputs with half-sized dimensions. Each convolution uses a kernel (or filter) of size 3 and is activated by a Leaky ReLU with a negative slope coefficient of 0.2. The contracting path extracts and learns the input's hierarchical features which are then diffused in the expanding path that tries to optimize alignment at successively finer resolutions by estimating ϕ.

Compared to VoxelMorph's proposed CNNs, We have used more layers and feature maps to cater for larger images and capture more features. Also, our overall network structure looks closer to the MICCAI's version of VM and similar at the bottom part — between the contracting and expanding blocks — to the original U-Net model. Furthermore, our network automatically

builds its structure given the number of layers defined by the user whereas the VM model is restricted to a fixed number of layers.

In our experiments, we have used images of size 512x448 but other dimensions (including depth) are possible. In the case of 3D images, the output of the CNN is of depth 3 where each channel is a volume storing the displacements in the different directions.

3.6. Training Strategy

We have organized the data into folders. Each folder contains a pair of mammogram images to be registered. The training was based on CC-view mammograms which make only 133 cases. Instead of splitting this small dataset into train and test sets, we did a 7-fold cross-validation (Figure 9). Specifically, the data was partitioned into 7 parts. One part served as the test set while the rest formed the training set. This step was repeated 7 times and at each cycle a different fold was fixed as the test set. At the end, all the 133 cases were registered and used for evaluation.

3.7. Evaluation

Assessing a registration result is not as straightforward as for other medical imaging tasks such as detection and classification. The ideal mean to measure our method's accuracy is to compare the obtained deformation fields with ground truth warp fields. However, in the lack of such data, we resolved on using alternative methods, both quantitative and qualitative, for evaluating the registration performance.

3.7.1. Mutual Information (MI)

While the MSE and LCC metrics can be reused to quantitatively estimate the degree to which the fixed image and registered image are similar, we decided to utilize a more flexible metric known as Mutual Information (Wells III et al., 1996). MI relies on image entropy to measure statistical dependence between images. A higher value indicates better alignment. MI has the advantage of not being affected when the same anatomical structure in two images has different intensities such as the case in multi-modal scans. The formula is shown is Equation 5.

$$MI\,(X, Y) = \sum_i \sum_j p_{xy}(i, j) log \frac{p_{xy}(i, j))}{p_x(i) p_y(j))} \qquad (5)$$

where:

- X and Y are random variables associated with the images to be compared,
- p_{xy} is the joint probability mass function of X and Y,
- p_x and p_y are respectively the marginal probability mass function of X and Y.

3.7.2. Temporal Subtraction Image

One way to visually assess the registration result is to look at the deformed image and compare it with the fixed image. However, this direct comparison can sometimes be delicate and is intrinsically subjective. One way to mitigate this complication is by building a subtraction image that shows only the differences between two images such that when the two images are identical, the subtraction image is a black image. The usefulness of this method goes further in that it allows radiologists to better detect subtle information by eliminating normal background breast tissue. For instance, the study conducted by Katsuragawa et al. (2002) showed that temporal subtraction images substantially improved radiologists' detection of lung nodules. This technique is also very useful to obtain vascular information of breast lesions in contrast-enhanced mammography (Carton et al., 2008; Dromain and Balleyguier, 2010).

3.7.3. Deformation Field

While similarity metrics and subtraction images give a quick clue on the alignment success of a registration, they cannot be solely relied upon for quality assurance. In fact, a perfect registration may result from unrealistic deformations that do not model physically possible elastic movements. Therefore, it is important to check the deformation field for irregularities which also helps in tuning the regularization parameter λ.

3.8. Implementation

All the coding was done in Python programming language to keep in line with the original VoxelMorph implementation. The network was developed using Keras, a popular open-source deep learning library (Chollet et al., 2015), and runs on top of TensorFlow (Abadi et al., 2016). VoxelMorph uses the Adam optimizer (Kingma and Ba, 2014); in our framework, we adopted RMSprop (Tieleman and Hinton, 2012) since our trials showed faster convergence and more stability during long epochs when using the latter algorithm. The learning rate was fixed to 10^{-4}.

3.9. Experiments

3.9.1. Registering CC-View Mammograms

In the following set of experiments, we registered all the CC-view images using the strategy described in 3.6. We sought to understand how different similarity metrics, regularization parameters, and CNN architecture affect the registration performance.

1. Similarity Metrics and Regularization Parameters: We have made four models of VoxelMorphMammography by training the network in turn with a high and low regularization parameter λ and for each of the similarity metrics. With LCC, we tested with $\lambda = 1$ and $\lambda = .01$ whereas with MSE we used $\lambda = .3$ and $\lambda = .03$.

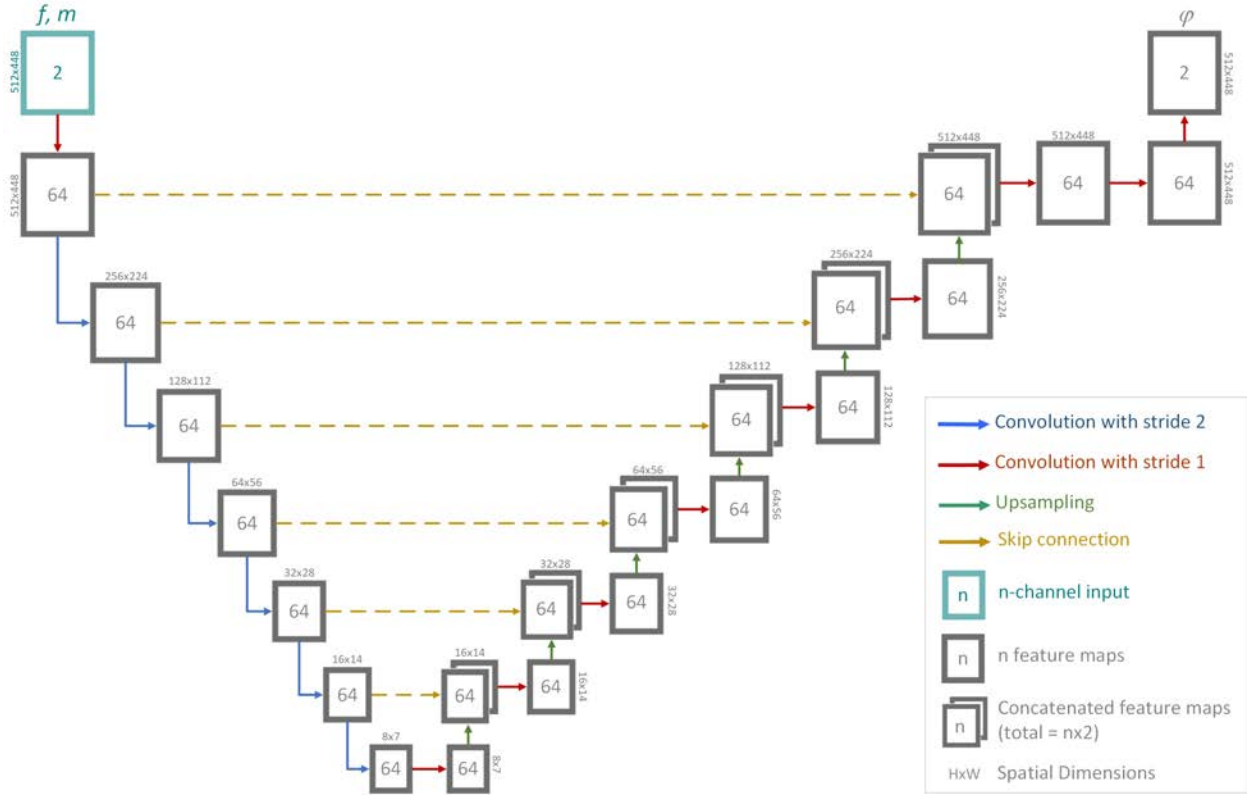


Figure 8: VoxelMorphMammography's CNN architecture

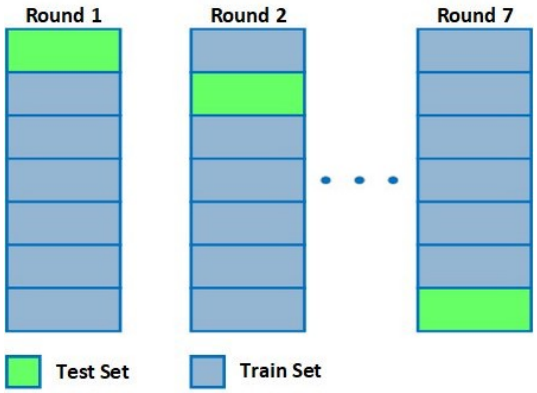


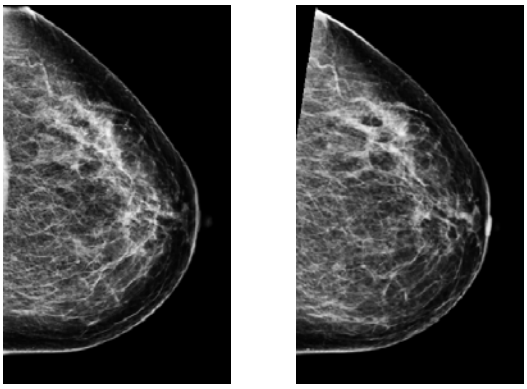
Figure 9: 7-Fold cross-validation

2. Non-diffeomorphic VM:
   In this experiment we wanted to test the effect of the network architecture and size to the alignment results. For this end, we carried the same previous four model trainings replacing the network with the original VoxelMorph CNN (Figure 6).

3. Diffeomorphic VM: Diffeomorphism preserves topological properties and produces invertible deformation fields. To compare the registration quality between diffeomorphic and non-diffeormophic formulations, we experimented with the more recent implementation of VoxelMorph that uses the CNN in Figure 7. This version uses a different and more complex loss function; its details are described in Dalca et al. (2019).

3.9.2. Registering MLO-View Mammograms

To test the performance of VoxelMorphMammography in registering mammograms taken at a view different from the view used to train it, we fixed the CC-view images as the training set and trained two models for 1000 epochs, one using the MSE metric and the other using the LCC metric. Each model was then used to register the MLO-view images.



(a) Fixed image (b) Moving image

Figure 10: Sample temporal mammogram pair. The moving image was affinely aligned to the fixed image during the data pre-processing step.

4. Results

4.1. VoxelMorphMammography Experiments

The initial experiments revolved around training our designed network with different parameter settings. We used both the MSE and LCC metrics for the loss function and in combination with each, we set with different values for the regularization parameter λ. The qualitative results (warped images, subtraction images and deformation fields) will be based on the fixed and moving image shown in Figure 10. We remind that the non-rigid transformations are applied on affinely aligned images. The training and testing of the neural networks involved the 133 pairs of CC-view mammograms.

4.1.1. Mutual Information Results

Figure 14 shows the boxplots of the computed mutual information for all the CC-view image pairs using the separately trained models. The plot also includes the MI values computed before any registration, after the initial global alignment using the affine transform, and after using the baseline method based on B-splines. We can see that the simplest form of transformation, i.e. affine, provides a good initial alignment of the temporal mammograms. The baseline method pushes the registration performance even higher but has longer whiskers indicating some pairs of images benefited from significant improvement in alignment than others. Its lower whisker, however, shows that some pairs experienced a drop in MI with the baseline method. The VMM models using a small regularization weight (0.01 for LCC and 0.03 for MSE) achieve higher MI values compared to the baseline. When using LCC, a higher regularization weight ($\lambda = 1$) yields a boxplot almost similar to the B-spline method. The model trained with MSE and $\lambda = 0.3$ gave the lowest MI scores among the non-rigid methods; yet, it has the same median as VMM trained with LCC and $\lambda = 1$. The figure includes the results of models trained with only two λ values, but in practice we have experimented with different values and obtained the same pattern: the lower the regularization penalty, the higher the MI, and vice-versa.

	Change of MI compared to affine		
Methods	**Decrease**	**Increase**	**No change**
Baseline (B-spline)	4	125	4
VMM (s=CC,l=1)	0	133	0
VMM (s=LCC,l=.01)	0	133	0
VMM (s=MSE,l=.3)	0	133	0
VMM (s=MSE,l=.03)	0	133	0

Table 1: Number of pairs that have seen a decrease, increase or no change in MI value using different deformable registration methods compared to the affine alignment.

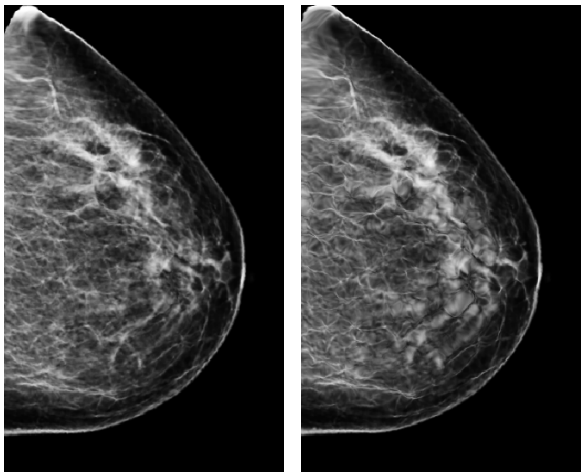
Table 1 helps to see how the MI value changed by using the different registration methods on affinely registered data. The baseline method resulted in a drop of MI for 4 pairs of mammograms and no significant difference for another 4 pairs. In contrast, all the deep learning-based methods improved the MI score over the whole dataset.

4.1.2. Temporal Subtraction Images

Figure 18 shows, for each method, the subtraction image obtained by taking the absolute difference of the fixed image and the warped image. Visually, these images can provide beneficial details as stated previously; nevertheless, we can provide a numerical value to quantify the amount of non-suppressed pixels as an error computed using the MSE equation. This value has been added under each subtraction image for facilitating comparison between the methods. VMM models using low regularization penalties produced the smallest errors. Only in combination with the MSE metric that a higher λ value gave an error lower than the baseline's. The model trained with a LCC-driven loss function and $\lambda = 1$ has a larger error; visually, the top-left corner in Figure 18b has a relatively larger white area, i.e. more differences between the subtracted images.

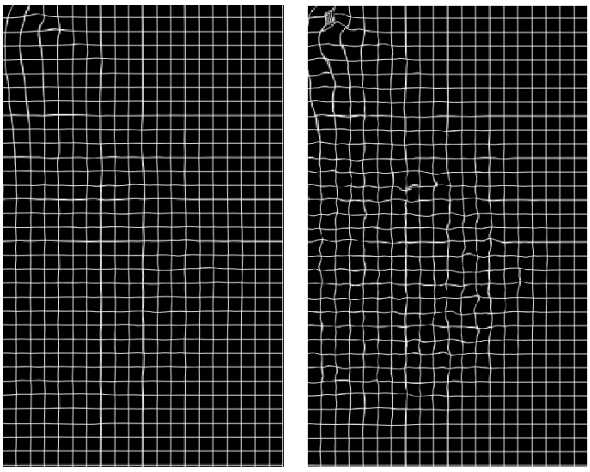
4.1.3. Deformation Fields

Warp fields can be displayed in a variety of ways. As an image, shades of primary colors (e.g. RGB) characterize the direction and intensity of displacements. Quiver plots are also commonly used for displaying these fields. Another method yet, is to warp an image of repeated vertical and horizontal lines, i.e. a grid, using the found deformation field. The deformed grid shows how points have moved position while its overall appearance gives an insight about the smoothness of the warp field. Figure 19 shows the warped grids we have obtained. We can notice high irregularities with lower λ's while the baseline method and the VMM model trained with MSE and a higher λ generates a clear and smooth deformed structure.



(a) $\lambda = 100$ (b) $\lambda = 10$

Figure 11: Deformed images obtained using the diffeomorphic formulation of VoxelMorph. The difference error (MSE) with the fixed image is 0.014 for (a) and 0.0074 for (b).



(a) $\lambda = 100$ (b) $\lambda = 10$

Figure 12: Deformation fields obtained using the diffeomorphic formulation of VoxelMorph.

4.2. VoxelMorph Experiments

Using the same dataset and set of parameters, we have trained new models using the original VoxelMorph network. The MI results are displayed in Figure 15. Overall, our architecture yielded better scores especially at lower values of the regularization parameter. The only case when VM improved on the VMM is with the MSE metric and with $\lambda = 0.3$, but the improvement is slight while the median MI is similar for both. Qualitative measures are not presented for these experiments as both frameworks generates very similar results with a small degree of perceived better visualization quality from using VMM.

4.3. Diffeomorphism Experiments

The VoxelMorph framework incorporating a diffeomorphic formulation has more complexity in parameters setting. It would require better understanding of the inner mechanism and more trials to achieve a fairer evaluation and comparison with the other models. Nevertheless, we show in Figure 16 the results of two trained models from using a high and low regularization weight along with the results previously obtained with VMM models. The diffeomorphic model achieved similar MI results to the VMM model with MSE metric when higher regularization weights were applied. Lowering the weight slightly improved the MI values but there are still behind the values obtained with VMM models. However, the diffeomorphic model at low λ produced the lowest difference error (0.0074). Figure 11 and Figure 12 contain, respectively, the warped images and deformation fields from these experiments. We can notice how the warp field maintains regularity to some degree even at reduced regularization penalty (Figure 12b).

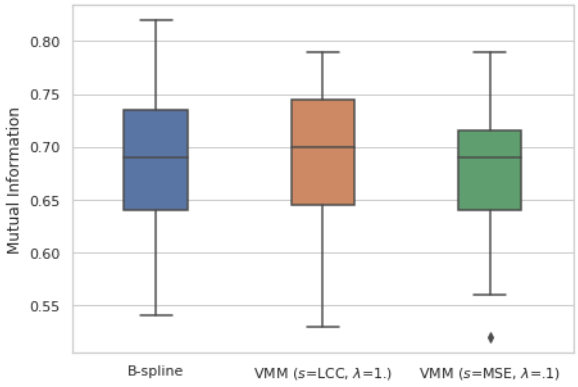


Figure 13: Boxplots of MI values for 27 pairs of temporal MLO-view mammograms registered using B-spline transform and a VoxelMorphMammography model trained on 130 pairs of CC-view images.

4.4. Registering MLO Mammograms

In this experiment, we focused on using a high regularization weight since antecedent tests always yielded optimized metrics with low λ values. For both LCC and MSE trials, we fixed this parameter to 0.1 and the obtained MI results are presented in Figure 13. The plot of VMM with LCC has higher MI median than the plot of the baseline method; yet, the latter has comparatively shorter lower whisker and longer upper whisper.

4.5. Execution Time

Post-training, VoxelMorphMammography takes on average 0.05s on the GPU to perform a single pairwise image registration. Using elastix running on a CPU, affine and B-spline registrations take, respectively, about 30s and 45s per case.

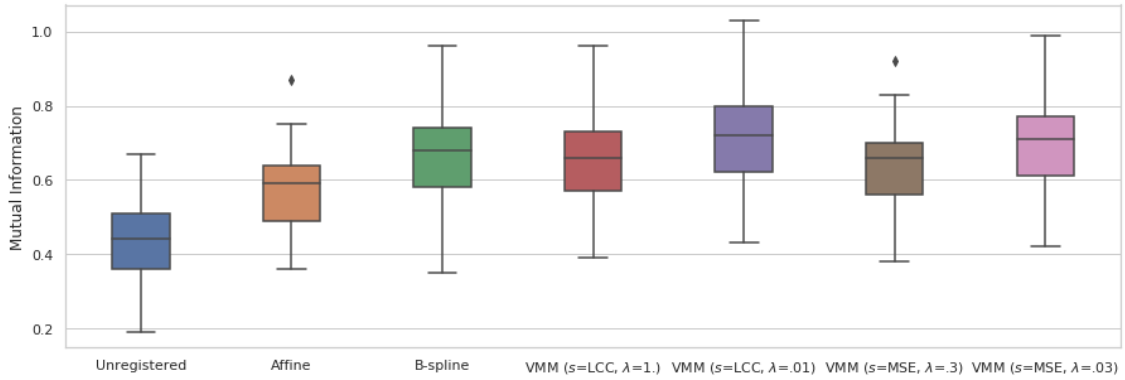


Figure 14: Boxplots of MI values for 130 pairs of temporal CC-view mammograms before registration and after registration using different methods. s denotes the metric adopted for the loss function while λ is the regularization weight.

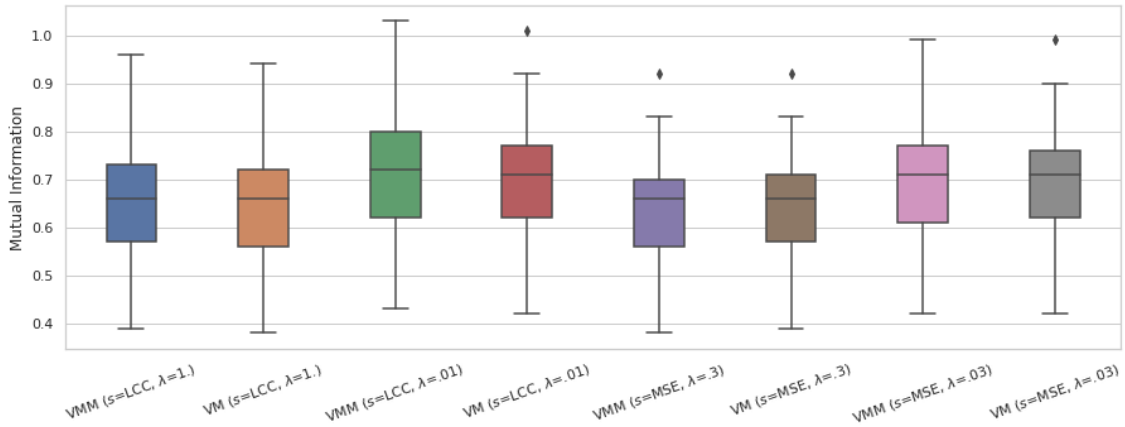


Figure 15: Boxplots of MI values for 130 pairs of temporal CC-view mammograms registered using VoxelMorphMammography and non-diffeomorphic VoxelMorph. The aim is to compare the performance of the 2 frameworks given the same combinations of parameters.

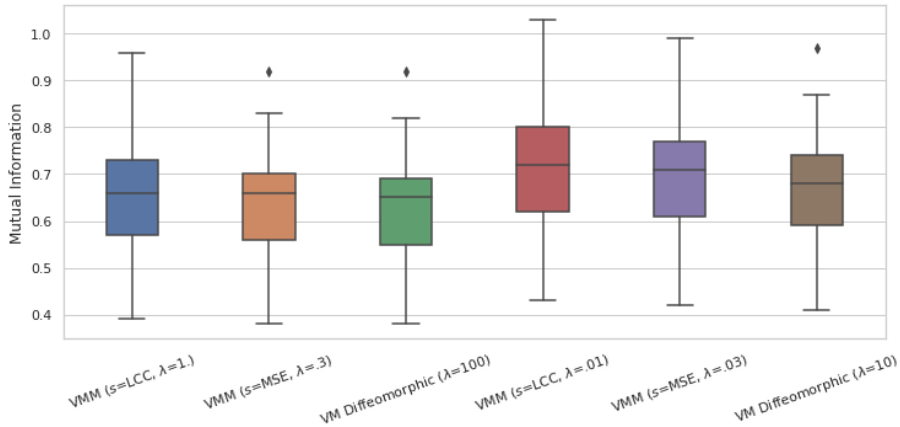
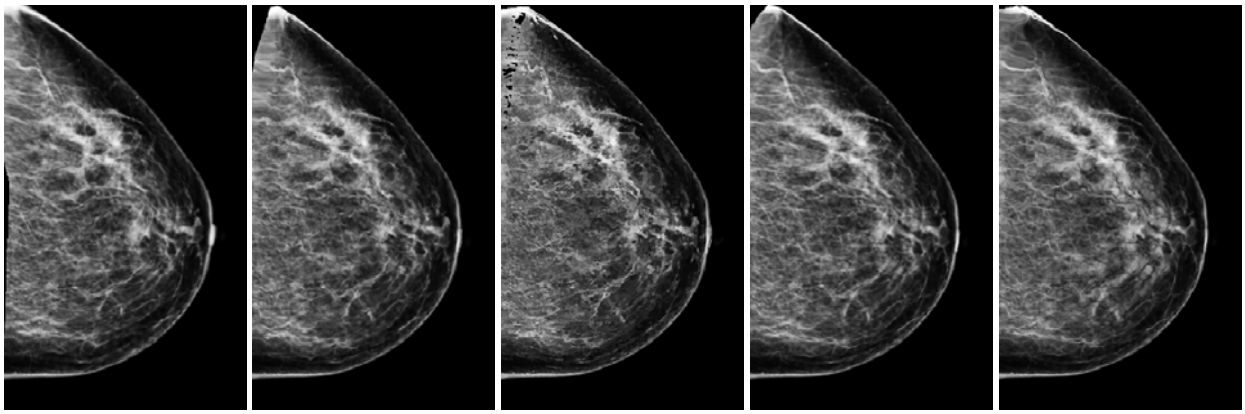
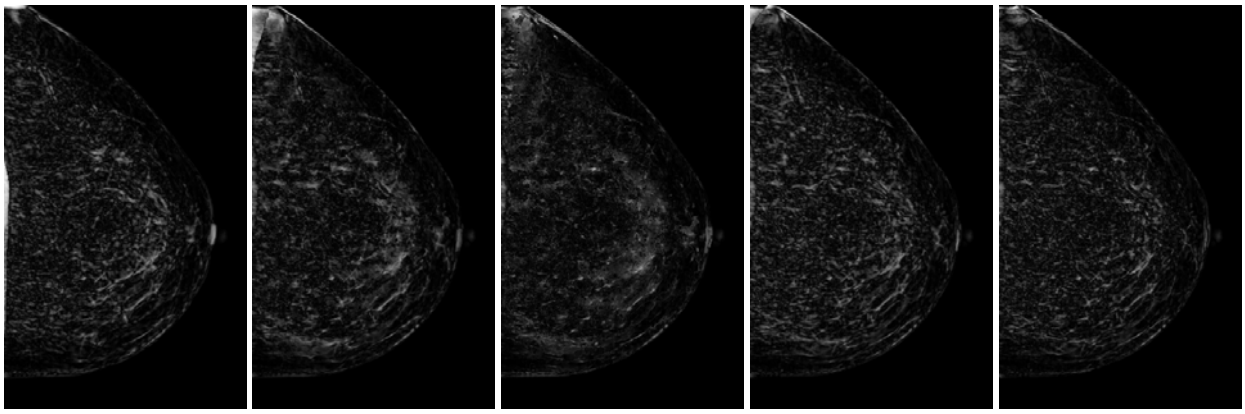


Figure 16: Boxplots of MI values for 130 pairs of temporal CC-view mammograms registered using VoxelMorphMammography and diffeomorphic VoxelMorph. The aim is to see the effect of diffeomorphism on the registration result.



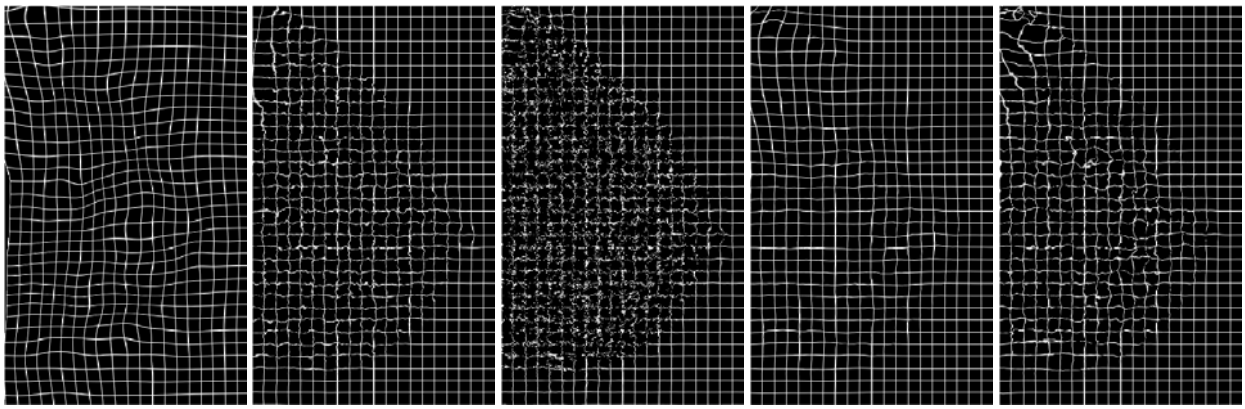
(a) Baseline (B-spline) MI=0.74 | (b) VMM, s=LCC, $\lambda = 1$ MI=0.75 | (c) VMM, s=LCC, $\lambda = .01$ MI=0.82 | (d) VMM, s=MSE, $\lambda = .3$ MI=0.72 | (e) VMM, s=MSE, $\lambda = .03$ MI=0.79

Figure 17: Results of registering the moving image (Figure 10b) using different methods. As a reference, the MI before registration is 0.62.



(a) Baseline (B-spline) error=0.0108 | (b) VMM, s=LCC, $\lambda = 1$ error=0.0111 | (c) VMM, s=LCC, $\lambda = .01$ error=0.0082 | (d) VMM, s=MSE, $\lambda = .3$ error=0.0105 | (e) VMM, s=MSE, $\lambda = .03$ error=0.0075

Figure 18: Subtraction images: difference between fixed image (Figure 10a) and warped images (Figure 17). The error quantifies the difference numerically and was computed using the MSE formula. As a reference, the error before registration is 0.0169.



(a) Baseline (B-spline) | (b) VMM, s=LCC, $\lambda = 1$ | (c) VMM, s=LCC, $\lambda = .01$ | (d) VMM, s=MSE, $\lambda = .3$ | (e) VMM, s=MSE, $\lambda = .03$

Figure 19: Deformation fields displayed as warped grids.

5. Discussion

We can notice that the hints given by the MI values about the degree of alignment are reflected to some extent in the subtracted images in the sense that more structures are suppressed in theses images when the MI between fixed and warped image is higher. However, this may not always be the case. For example, consider the case of registering the image pair in Figure (10), the VMM model with MSE as metric for the loss function and lower λ value gave the second highest MI (Figure 17e), but when considering the MSE value (Figure 18e) this model gave the lowest error. Conversely, the model using the LCC metric and lower regularization weight generates the highest MI (Figure 17c) but gives the second lowest error (Figure 18c). Deciding on which of these two models is more successful at registering the sample image pair needs other indicators to be considered. If we analyze their corresponding registration fields (Figure 19), we see that the first model produces much less warp irregularities than the second; therefore, the model with the MSE metric would be a relatively better model to adopt.

Overall, using the MSE metric in the loss function instead of LCC produced better qualitative results for the type of images used. During the training phase, we saw that the training loss converges quickly with MSE while the LCC continuously minimizes the error which explains the tendency of the LCC models to overfit to the fixed image. Maintaining a higher regularization weight when using LCC is therefore necessary.

Furthermore, we notice that using a larger CNN architecture provides better registration results. As for the diffeomorphic models, they effectively ensured reasonably smooth warp fields even though the MI values were relatively low. More experiments on these models and their parameters will be a good idea for future work.

Finally, we can confidently train VoxelMorphMammography with CC and MLO mammograms at the same time. Eventually, training with multiple databases could also be possible and may yield a more generic model for mammogram image registration.

6. Conclusions

In this paper, we have presented VoxelMorphMammography (VMM), an extension of the VoxelMorph framework tailored for the registration task of temporal mammograms. VMM is based on convolutional neural networks and spatial transformers and can be deployed for registering 3D volumes even though in our work we restricted the experiments to registering 2D mammograms. VMM is also a self-learning neural network and therefore ground truth data is not required. Compared with a state-of-the-art toolbox for medical image registration — elastix, VMM trained with only 133 mammogram pairs can achieve comparable results while requiring a fraction of second to perform pairwise image registration. This is made possible by optimizing a CNN-modeled global function during training instead of solving for each image pair. The optimization of the registration performance relies on a loss function that couples a similarity metric, such as cross-correlation and mutual square error, with a diffusion regularizer that penalizes for irregular spatial variations in the deformation field. Additionally, VMM has demonstrated its capability to register mammograms taken at a projection different from the one used during training of the network. This study comes at the conclusion that deep learning approaches have the potential to replace traditional methods for mammograms registration. However, there is a need for conducting more research and experimentation to provide quality assurance through establishing a robust validation method.

7. Acknowledgments

This work has been conducted at the University of Girona in partial fulfillment of the requirements for the Master's degree in Medical Imaging and Applications (MAIA) under the supervision of Dr. Robert Marti and Dr. Oliver Diaz whose guidance and suggestions were helpful and meaningful.

Special thanks are extended to Adrian Vasile Dalca, postdoctoral fellow at Massachusetts Institute of Technology and co-author of the VoxelMorph paper for his help in clarifying doubts revolving around VoxelMorph and his suggestions for improving the results.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al., 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467 .

Abdel-Nasser, M., Moreno, A., Puig, D., 2016. Temporal mammogram image registration using optimized curvilinear coordinates. Computer methods and programs in biomedicine 127, 1–14.

Bağcı, U., Udupa, J.K., Bai, L., 2010. The role of intensity standardization in medical image registration. Pattern Recognition Letters 31, 315–323.

Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2018. An unsupervised learning model for deformable medical image registration, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 9252–9260.

Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2019. Voxelmorph: a learning framework for deformable medical image registration. IEEE transactions on medical imaging .

Bassett, L.W., Shayestehfar, B., Hirbawi, I., 1994. Obtaining previous mammograms for comparison: usefulness and costs. AJR. American journal of roentgenology 163, 1083–1086.

Bodhwani, V., Acharjya, D., Bodhwani, U., 2019. Deep residual networks for plant identification. Procedia Computer Science 152, 186–194.

Boehler, T., Zoehrer, F., Harz, M., Hahn, H.K., 2012. Breast image registration and deformation modeling. Critical Reviews¢ in Biomedical Engineering 40.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A., 2018. Global cancer statistics 2018: Globocan estimates

of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians 68, 394–424.

Cao, X., Yang, J., Zhang, J., Nie, D., Kim, M., Wang, Q., Shen, D., 2017. Deformable image registration based on similarity-steered cnn regression, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 300–308.

Carneiro, T., Da Nóbrega, R.V.M., Nepomuceno, T., Bian, G.B., De Albuquerque, V.H.C., Reboucas Filho, P.P., 2018. Performance analysis of google colaboratory as a tool for accelerating deep learning applications. IEEE Access 6, 61677–61685.

Carton, A.K., Currivan, J.A., Conant, E., Maidment, A., 2008. Temporal subtraction versus dual-energy contrast-enhanced digital breast tomosynthesis: A pilot study, in: International Workshop on Digital Mammography, Springer. pp. 166–173.

Chollet, F., et al., 2015. Keras. URL: `https://github.com/fchollet/keras`.

Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R., 2019. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. arXiv preprint arXiv:1903.03545 .

Diez, Y., Oliver, A., Lladó, X., Freixenet, J., Marti, J., Vilanova, J.C., Marti, R., 2011. Revisiting intensity-based image registration applied to mammography. IEEE Transactions on Information Technology in Biomedicine 15, 716–725.

Dromain, C., Balleyguier, C., 2010. Contrast-enhanced digital mammography, in: Digital mammography. Springer, pp. 187–198.

Ekeh, A.P., Alleyne, R.S., Duncan, A.O., 2000. Role of mammography in diagnosis of breast cancer in an inner-city hospital. Journal of the National Medical Association 92, 372.

van Engeland, S., Snoeren, P.R., Karssemeijer, N., Hendriks, J.H., 2003. Optimized perception of lesion growth in mammograms using digital display, in: Medical Imaging 2003: Image Perception, Observer Performance, and Technology Assessment, International Society for Optics and Photonics. pp. 25–32.

Fitzmaurice, C., Allen, C., Barber, R.M., Barregard, L., Bhutta, Z.A., Brenner, H., Dicker, D.J., Chimed-Orchir, O., Dandona, R., Dandona, L., et al., 2017. Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the global burden of disease study. JAMA oncology 3, 524–548.

Galceran, J., Ameijide, A., Carulla, M., Mateos, A., Quirós, J., Rojas, D., Alemán, A., Torrella, A., Chico, M., Vicente, M., et al., 2017. Cancer incidence in spain, 2015. Clinical and Translational Oncology 19, 799–825.

Google, 2019. Welcome to colaboratory. URL: `https://colab.research.google.com`. accessed: 2019-06-06.

Hipwell, J.H., Vavourakis, V., Han, L., Mertzanidou, T., Eiben, B., Hawkes, D.J., 2016. A review of biomechanically informed breast image registration. Physics in Medicine & Biology 61, R1.

Jaderberg, M., Simonyan, K., Zisserman, A., et al., 2015. Spatial transformer networks, in: Advances in neural information processing systems, pp. 2017–2025.

Katsuragawa, S., Uozumi, T., Kakeda, S., Watanabe, H., Nakata, H., Doi, K., 2002. Clinical usefulness of temporal subtraction technique for detection of interval changes on digital chest radiographs, in: CARS 2002 Computer Assisted Radiology and Surgery. Springer, pp. 689–694.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Klein, A., Andersson, J., Ardekani, B.A., Ashburner, J., Avants, B., Chiang, M.C., Christensen, G.E., Collins, D.L., Gee, J., Hellier, P., et al., 2009. Evaluation of 14 nonlinear deformation algorithms applied to human brain mri registration. Neuroimage 46, 786–802.

Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P., 2010. Elastix: a toolbox for intensity-based medical image registration. IEEE transactions on medical imaging 29, 196–205.

Krebs, J., Mansi, T., Delingette, H., Zhang, L., Ghesu, F.C., Miao, S., Maier, A.K., Ayache, N., Liao, R., Kamen, A., 2017. Robust non-rigid registration through agent-based action learning, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 344–352.

Kwok, S.M., Chandrasekhar, R., Attikiouzel, Y., Rickard, M.T., 2004. Automatic pectoral muscle segmentation on mediolateral oblique view mammograms. IEEE transactions on medical imaging 23, 1129–1140.

Li, H., Fan, Y., 2017. Non-rigid image registration using fully convolutional networks with deep self-supervision. arXiv preprint arXiv:1709.00799 .

Maas, A.L., Hannun, A.Y., Ng, A.Y., 2013. Rectifier nonlinearities improve neural network acoustic models, in: Proc. icml, p. 3.

Marias, K., Behrenbruch, C., Parbhoo, S., Seifalian, A., Brady, M., 2005. A registration framework for the comparison of mammogram sequences. IEEE Transactions on Medical Imaging 24, 782–790.

Marias, K., Brady, J., Highnam, R., Parbhoo, S., Seifalian, A., Wirth, M., 1999. Registration and matching of temporal mammograms for detecting abnormalities. Medical Imaging Understanding and Analysis .

Mertzanidou, T., Hipwell, J., Cardoso, M.J., Zhang, X., Tanner, C., Ourselin, S., Bick, U., Huisman, H., Karssemeijer, N., Hawkes, D., 2012. Mri to x-ray mammography registration using a volume-preserving affine transformation. Medical image analysis 16, 966–975.

Modat, M., Ridgway, G.R., Taylor, Z.A., Lehmann, M., Barnes, J., Hawkes, D.J., Fox, N.C., Ourselin, S., 2010. Fast free-form deformation using graphics processing units. Computer methods and programs in biomedicine 98, 278–284.

Pinto Pereira, S.M., Hipwell, J.H., McCormack, V.A., Tanner, C., Moss, S.M., Wilkinson, L.S., Khoo, L.A., Pagliari, C., Skippage, P.L., Kliger, C.J., et al., 2010. Automated registration of diagnostic to prediagnostic x-ray mammograms: Evaluation and comparison to radiologists accuracy. Medical physics 37, 4530–4539.

Richard, F.J., Bakic, P.R., Maidment, A.D., 2006. Mammogram registration: a phantom-based evaluation of compressed breast thickness variation effects. IEEE transactions on medical imaging 25, 188–197.

Rohlfing, T., Maurer, C.R., Bluemke, D.A., Jacobs, M.A., 2003. Volume-preserving nonrigid registration of mr breast images using free-form deformation with an incompressibility constraint. IEEE transactions on medical imaging 22, 730–741.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

Roshni, V., Revathy, K., 2008. Using mutual information and cross correlation as metrics for registration of images. Journal of Theoretical & Applied Information Technology 4.

Roy, S., Panda, A., Naskar, R., 2019. Unsupervised ground truth generation for automated brain em image segmentation, in: 2019 6th International Conference on Signal Processing and Integrated Networks (SPIN), IEEE. pp. 66–71.

Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L., Leach, M.O., Hawkes, D.J., 1999. Nonrigid registration using free-form deformations: application to breast mr images. IEEE transactions on medical imaging 18, 712–721.

Satılmış, Y., Tufan, F., Şara, M., Karslı, M., Eken, S., Sayar, A., 2018. Cnn based traffic sign recognition for mini autonomous vehicles, in: International Conference on Information Systems Architecture and Technology, Springer. pp. 85–94.

Schopper, D., de Wolf, C., 2009. How effective are breast cancer screening programmes by mammography? review of the current evidence. European journal of cancer 45, 1916–1923.

Sokooti, H., de Vos, B., Berendsen, F., Lelieveldt, B.P., Išgum, I., Staring, M., 2017. Nonrigid image registration using multi-scale 3d convolutional neural networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 232–239.

Sree, S.V., Ng, E.Y.K., Acharya, R.U., Faust, O., 2011. Breast imaging: a survey. World journal of clinical oncology 2, 171.

Tieleman, T., Hinton, G., 2012. Lecture 6.5-rmsprop: Divide the gra-

dient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning 4, 26–31.

Timp, S., van Engeland, S., Karssemeijer, N., 2005. A regional registration method to find corresponding mass lesions in temporal mammogram pairs. Medical physics 32, 2629–2638.

Tortajada, M., Oliver, A., Martí, R., Ganau, S., Tortajada, L., Sentís, M., Freixenet, J., Zwiggelaar, R., 2014. Breast peripheral area correction in digital mammograms. Computers in biology and medicine 50, 32–40.

Tortajada, M., Oliver, A., Martí, R., Vilagran, M., Ganau, S., Tortajada, L., Sentís, M., Freixenet, J., 2012. Adapting breast density classification from digitized to full-field digital mammograms, in: International Workshop on Digital Mammography, Springer. pp. 561–568.

de Vos, B.D., Berendsen, F.F., Viergever, M.A., Staring, M., Išgum, I., 2017. End-to-end unsupervised deformable image registration with a convolutional neural network, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 204–212.

Wai, L.C., Brady, M., 2005. Curvilinear structure based mammographic registration, in: International Workshop on Computer Vision for Biomedical Image Applications, Springer. pp. 261–270.

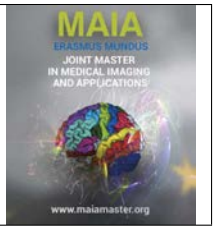
Wei, C.H., Gwo, C.Y., Huang, P.J., 2016. Identification and segmentation of obscure pectoral muscle in mediolateral oblique mammograms. The British journal of radiology 89, 20150802.

Wells III, W.M., Viola, P., Atsumi, H., Nakajima, S., Kikinis, R., 1996. Multi-modal volume registration by maximization of mutual information. Medical image analysis 1, 35–51.

Yang, X., Kwitt, R., Styner, M., Niethammer, M., 2017. Quicksilver: Fast predictive image registration–a deep learning approach. NeuroImage 158, 378–396.

Medical Imaging and Applications

Master Thesis, June 2019



Clinical Outcome Prediction in Acute Ischemic Stroke using Traditional Machine Learning and Convolutional Neural Networks

Márcio Aloísio Bezerra Cavalcanti Rockenbach, Arnau Oliver, Mariano Cabezas, Xavier Lladó

University of Girona, Computer Vision and Robotics Institute, Edifici P-IV. Campus de Montilivi, 17003, Girona (Spain)

Abstract

Acute ischemic stroke is a highly prevalent condition that can lead to a significant impairment in patient's life. The development of techniques that can anticipate clinical outcome has the potential to aid physicians in properly addressing patient management. This project developed strategies to predict clinical outcome from magnetic resonance (MR) images and from clinical data. Different strategies were implemented in terms of 1) preprocessing steps, 2) feature extraction approach (traditional machine learning versus feature extraction using convolutional neural networks), 3) input to the model (regions of interest based on lesion segmentation versus whole brain volume) and 4) classification parameters (types of classifiers, dimensionality reduction techniques and clinical information usage). The goal was to predict the modified Rankin Scale score, which measures the degree of disability after a stroke event by assigning a score that varies from zero to six. The dataset was provided by the ISLES 2016 Challenge, including 30 cases. The metric used to evaluate the performance of the models was the mean absolute error (MAE) between the computed and the actual scores. The best result from the traditional machine learning strategy was obtained by extracting whole volume features from a 3D image generated from the subtraction of two perfusion MR sequences, namely cerebral blood volume (CBV) and cerebral blood flow (CBF), which achieved a MAE of 0.43 ± 0.76. For the convolutional neural network approach, feature extraction was performed using the encoder part of a 3D U-net, which was pre-trained to perform the lesion segmentation of images from the same dataset. The best result was obtained using the whole CBF/CBV difference volume as network input, leading to MAE of 0.50 ± 0.95. Approaches that did not require lesion segmentation achieved similar or even better results than strategies that used it, demonstrating how radiological knowledge can be incorporated into project design to guide extraction of the most meaningful information from the available resources. Future work with larger datasets is necessary to allow a more robust evaluation of the generated models.

Keywords: stroke, clinical outcome, clinical prediction, MRI, perfusion, ISLES, medical challenges, machine learning, convolutional neural networks

1. Introduction

Stroke is defined as an acute episode of focal dysfunction of the brain, retina, or spinal cord lasting longer than 24h, or of any duration if imaging such as computed tomography (CT) or magnetic resonance (MR) or if autopsy show focal infarction or hemorrhage relevant to the symptoms (Hankey, 2017). The most typical symptoms of stroke include sudden unilateral weakness, numbness, visual loss, diplopia, altered speech, ataxia and non-orthostatic vertigo (Hankey and Blacker, 2015).

According to the American Heart Association (AHA), stroke prevalence in adults is 2.7º% in the United States (Benjamin et al., 2018). Each year, around 795,000 people experience a new or recurrent stroke. In Spain, 116,017 cases were reported in 2013 (Brea et al., 2013). Regarding the type, 87% were ischemic and 13% were hemorraghic (being 10% intracerebral hemorraghic and 3% subarachnoid hemorrhagic). This condition represents also a tremendous social burden, not only by the mortality but also by the associated morbidity (Vilela and Rowley, 2017). As a consequence of

population ageing, it is expected that the health costs will continue to rise exponentially over the next decades (Favate and Younger, 2016).

1.1. Management of Ischemic Acute Stroke

For many years, treatment of acute ischemic stroke was based only on supportive therapy, rehabilitation and risk management. Nowadays, guidelines recommend that patients within 4.5 hours of symptom onset should be treated with intravenous tissue plasminogen activator (tPA), if there are no contraindications (Meschia and Brott, 2018). The goal of using tPA is to dissolve the clot that caused the stroke, allowing the restoration of cerebral blood perfusion.

Stroke imaging already plays a key role in the decision making process regarding the choice of the correct treatment, being incorporated in guidelines for patient management. Several machine and deep learning approaches were already developed and implemented to forecast the outcome of the stroke lesion itself, but few of them dealt with clinical prediction. The development of strategies to predict final clinical patient outcome from imaging and non-imaging data have potential to provide new valuable information to guide stroke management, improving patient care even further.

1.2. Project Description and Goal

This master thesis focused on using MR images and clinical data provided by the Ischemic Stroke Lesion Segmentation (ISLES) Challenge 2016 to predict patient's clinical outcome in acute ischemic stroke.

Two main approaches were considered for feature extraction: traditional machine learning (ML) using hand-crafted features and convolutional neural networks (CNN). For both cases, effect of using features from regions of interest (ROIs), determined by lesion segmentation, or features from the whole brain volume (without lesion segmentation) were analyzed. Additionally, the effect of implementing different strategies in terms of preprocessing steps, classifiers, usage of clinical information and dimensionality reduction were also assessed.

The clinical outcome was measured through the modified Rankin Scale, which measures the degree of disability after a stroke event by giving a score that varies from zero to six. The experiment results were quantitatively evaluated on the training set of the challenge according to the mean absolute error (MAE) between the predicted labels and the true labels.

2. Medical Background

2.1. Imaging in Acute Stroke

Imaging plays a fundamental role in the diagnosis and in the management of stroke patients. The 2018 AHA Guidelines recommend that all patients admitted to hospital with suspected acute stroke should receive brain imaging evaluation upon hospital admission (Powers et al., 2018). In most cases, noncontrast CT (NCCT) provides the necessary information to make decisions about acute management. The guidelines also determine that selected patients may benefit from obtaining computed tomography perfusion (CTP), diffusion-weighted imaging (DWI), or perfusion-weighted imaging (PWI) to aid in patient selection for mechanical thrombectomy.

NCCT has limited sensitivity for the diagnosis of ischemic stroke during the initial hours, and it is necessary to improve the diagnostic accuracy to recommend optimal thrombolytic and other stroke therapies (Patel et al., 2001). Other MRI techniques such as DWI and PWI play a fundamental role in this matter, having the potential to improve the diagnosis while being practical and feasible.

DWI measures the net movement of water in tissue caused by random (Brownian) molecular motion of water and shows hyperintense ischemic tissue changes within minutes to a few hours after arterial occlusion due to a reduction of the apparent diffusion coefficient (ADC). The ADC reduction occurs primarily in the intracellular space associated with disruption in membrane ionic homeostasis and cytotoxic edema. Decrease signal in the ADC and increased signal on DWI studies represent irreversible ischemia (known as stroke core region). To differentiate acute from subacute or older lesions (effect known as T2 shine-through), DWI is used combined with T2-weighted images and ADC maps.

2.2. Perfusion MRI

Perfusion-weighted imaging (PWI) allows the measurement of capillary perfusion. The method most commonly used in clinical practice and in research is the dynamic susceptibility contrast-enhanced technique, in which paramagnetic contrast agent is injected as an intravenous (IV) bolus and the signal change is tracked by MR sequences (Edlow Jonathan et al., 2011). From the acquired data, we can derive several measurements (Allmendinger et al., 2012).

- Cerebral blood volume (CBV): measurement of the total volume of blood within an imaging voxel. Measured in units of mililiters of blood per 100g of brain.
- Cerebral blood flow (CBF): total volume of blood moving through a voxel in a given unit of time. Measured in units of mililiters of blood per 100g of brain tissue per minute.
- Mean transit time (MTT): average transit time of all the molecules of contrast medium with the bolus through a given volume of brain, measured in seconds. Can be approximated through the equation $MTT = CBV/CBF$.

- Time to peak enhancement (TTP): time from the start of the contrast injection to maximal enhancement, measured in seconds.
- Time to maximum (Tmax): time to maximum of the residue function obtained by deconvolution. The tissue contrast agent concentration C(t) can be expressed as a convolution of the arterial input function (AIF) and the residue function R(t): $C(t) = CBF \times (AIF(t) \circledast R(t))$. The residue function is obtained by deconvolution and its maximum value occurs, by definition, at Tmax (Calamante et al., 2010).

2.3. *Perfusion MRI in Acute Ischemic Stroke*

In acute ischemic stroke setting, perfusion MRI can be used to make the distinction between two main affected areas: core and penumbra. The core represents a area in the brain that has tissue with irreversible lesion, even if recanalization of the occluded artery is achieved early. The penumbra surrounds the core and is composed of salvageable brain tissue that might be recovered if recanalization of the occluded artery occurs promptly (Muir et al., 2006).

In terms of PWI parameters, the core is characterized by CBV and CBF decrease and MTT increase. It is still unsure which is the best parameter to define the infarct core, but there are evidences that CBF reduction (greater than 30% of the normal CBF) may have better correlation with DWI findings (Levi et al., 2011). In the penumbra region, we can expect increase in MTT and Tmax and decrease in CBF, with a relative preservation of the CBV. Therefore, mismatched areas of abnormal perfusion (prolonged MTT and diminished CBF where CBV is relatively preserved) are those that most likely correspond to penumbra.

2.4. *Modified Rankin Scale*

The Modified Rankin Scale (mRS) is commonly used for measuring the degree of disability or dependence in the daily activities of stroke patients. It was originally published by Rankin (1957), with scores ranging from one to six, with increasing degrees of disability. It was later modified by van Swieten et al. (1988) to add category zero, which represents the patients with no new disabilities after the event. Detailed description of each category can be seen in Table 1.

This scale is easy to use and has good inter-rater reliability, especially when using structured interviews (Bone et al., 2002) and is used to guide patient treatment. The 2018 Guidelines for Management of Acute Ischemic Stroke from the AHA (Powers et al., 2018) state that patients should receive mechanical thrombectomy with a stent retriever if they meet several criteria, which include having a prestroke modified Ranking Score (mRS) score of zero to one. This score is also widely utilized as a patient outcome criteria for several clinical trials (Barow et al., 2019; Uchida et al., 2019), being commonly assessed 90 days after the event.

Table 1: Description of the Modified Rankin Scale (mRS) categories.

Modified Rankin Scale (mRS)	
Category	**Description**
0	No symptoms at all.
1	No significant disability despite symptoms; able to carry out all usual duties and activities.
2	Slight disability; unable to carry out all previous activities, but able to look after own affairs without assistance.
3	Moderate disability; requiring some help, but able to walk without assistance.
4	Moderately severe disability; unable to walk without assistance and unable to attend to own bodily needs without assistance.
5	Severe disability; bedridden, incontinent and requiring constant nursing care and attention.
6	Dead.

2.5. *ISLES Challenges*

The ISLES is a medical image segmentation challenge that was held annually between 2015 and 2018 at the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI).

The ISLES Challenge 2016 (Egger et al., 2016) is currently the only one to deal not just with stroke lesion segmentation but also with clinical outcome prediction, and it was therefore chosen to be the focus of this work. For that year's challenge, the organizers provided MRI scans of acute stroke cases and their associated clinical parameters. The participants had to perform two tasks: lesion outcome (segmentation) and clinical outcome predictions. The associated ground truth was the final lesion volume (Task I) as manually segmented in 3 to 9 month follow-up scans, and the clinical mRS score (Task II) denoting the degree of disability was the final parameter to evaluate the clinical outcome. The overview of the challenge is shown in Figure 1.

3. State of the art

Most of the existing methods for clinical outcome prediction use clinical data to achieve the goal. Kabir et al. (2017) used a M5 tree model to predict the same mRS score used in this project. The selection of relevant characteristics from the electronic health record (EHR) was done by a neurologist specialized in stroke, resulting in 23 variables, among categorical and quantitative ones. Data was obtained from 439 patients, and they achieved a MAE of 0.54 in a 10-fold cross validation method.

Vrtkova (2017) also evaluated the mRS score from data obtained from the EHR, using 12 categorical and 5 quantitative variables from 449 patients. They modified

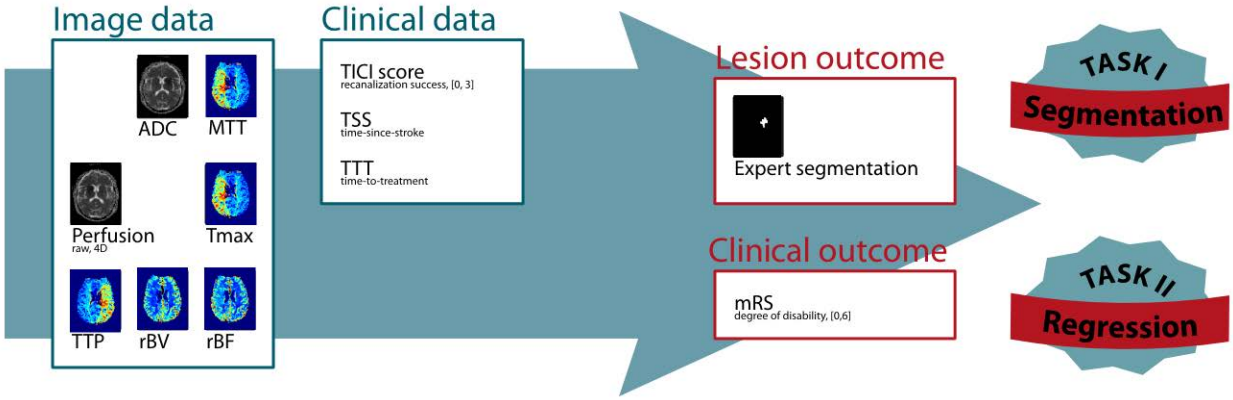


Figure 1. Ischemic Stroke Lesion Segmentation (ISLES) Challenge 2016 (Egger et al., 2016).

the task into a binary classification problem, by considering mRS categories zero, one, two and three in one group and scores four, five and six in another. A random forest approach reached 86% accuracy.

Bacchi et al. (2019) developed a pilot study to predict positive outcomes in stroke patients that underwent intravenous thrombolysis treatment using a deep learning approach. They used clinical data in combination with NCCT of 204 patients. Their goal was to predict which patients would have a mRS score of zero or one (which they considered as positive outcomes). They achieved an accuracy of 74%.

Clinical outcome prediction from MRI images is more scarce, and the current available approaches originated mostly from the ISLES Challenge 2016. Mahmood and Basit (2016) and Choi et al. (2016) used lesion segmentation in order to perform the prediction task. The current state-of-the-art method for this specific problem is described in Maier and Handels (2016), which achieved the best overall result for the second task (clinical prediction outcome) in this challenge. They achieved a MAE of 1.05 ± 0.62. The authors initially perform the lesion segmentation (Task I) using a random forest (RF) classifier. From the resulting segmentation, they implement a random regression forest that analyzes the obtained features to predict the final mRS score. In the next subsections, detailed description of this work is given.

3.1. Features for the Lesion Segmentation Task

To perform the lesion segmentation, the authors (Maier and Handels, 2016) created a RF classifier that was fed with different features extracted from the ADC and from the five perfusion maps (CBV, CBF, MTT, TTP, Tmax):

- Intensity features: voxel's unprocessed intensity value; voxel's intensity values after applying Gaussian smoothing in areas of 3, 5 and 7 mm around each voxel; intensity difference between corresponding voxels of the two brain hemispheres
- Distance features: 2D centerdistance (Euclidean distance to the central pixel of the slice) of each voxel (computed once for each of the three dimensions)
- Local histogram features: provides information about local intensity distribution in a small area around each voxel

The classifier was trained with 200 trees, producing a posteriori class probability map, which was thresholded at a value of 0.3 to create a binary image. The final post-processing step was the closing of structural 3D holes that resulted in the final binary segmentation mask.

3.2. Features for the Clinical Prediction Task

To perform the clinical prediction, the authors decided to extract features from three areas: (1) the lesion itself; (2) a band around the lesion whose fate was unclear; (3) the remainder of the brain. These three regions of interest (ROIs) were defined according to the binary mask returned by the lesion outcome prediction method. For this task, the probability map resulting from the first task was thresholded at a value of 0.1 instead of 0.3. To create the band around the lesion, the inner region mask was extended with a binary dilation of 5 mm. Finally, the remaining brain constituted the third region. This process is shown in Figure 2.

The next step was feature extraction. For this purpose, only images from the ADC sequence were used. The authors used a modified version of the same set of features implemented for the first lesion outcome, since those are voxel-wise image characteristics and not region based. To overcome this issue, they considered statistics of the feature values in each of the three regions: ten percentile values, standard deviation, variance and a histogram of ten bins. Geometric features of the three ROIs were also included: region area, perimeter, roundness, and equivdiameter (which corresponds to the diameter of a circle with the same area as the region).

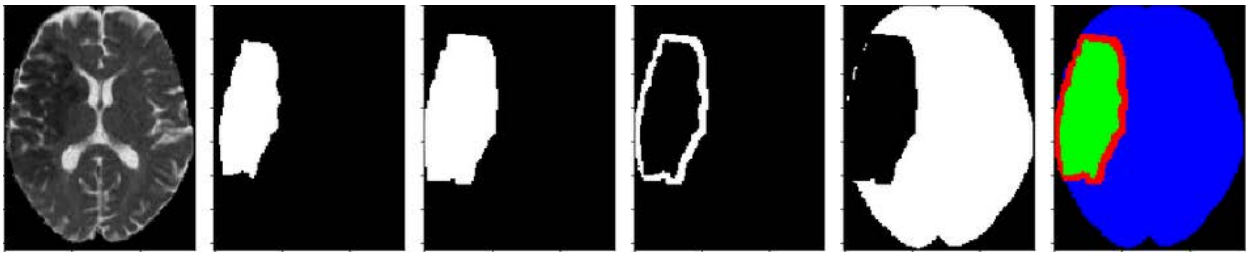


Figure 2. Segmentation of brain volume in three regions: stroke lesion (green), band around lesion (red) and rest of the brain (blue). From left to right, we have: ADC image, lesion ground truth/segmentation mask, dilation of the ground truth/segmentation mask, subtraction of the lesion from the dilation to create the band mask, rest of the brain mask and final regions of interest.

During the training phase, the authors made experiments incorporating the clinical information into the classifier, but it did not improve the results. For that reason, they decided not to use the provided clinical data in the final model. Finally, a RF classifier with 200 trees was trained using the mRS score as label. In the test set of the Challenge, the authors achieved a MAE of 1.05 ± 0.62, obtaining the first position.

4. Material and methods

4.1. Dataset

4.1.1. MRI Images

The challenge provided 30 MRI sequences for training and 19 for testing. All the available images were skull-stripped, anonymized and co-registered. The sequences included for each patient were ADC, PWI raw data and PWI perfusion maps (MTT, TTP, CBV, CBF and Tmax). The training set also contained binary volumes with the ground truth related to the lesion segmentation task. An example of the sequences provided for each patient is shown in Figure 3.

The images were obtained from different scanners and using different protocols. Some volumes cover the entire brain and the cerebellum, while some others contain less slices, mainly from the stroke region. By analyzing the dataset, it is possible to infer that the cases were acquired using four machines, resulting in images with the following characteristics:

- Scanner 1: 10 samples from training and 11 samples from testing; dimensions: 192 x 192 x 19 pixels; voxel spacing: 1.198 x 1.198 x 6.5 mm
- Scanner 2: 9 samples from training and 4 from testing; dimensions: 256 x 256 x 24 pixels; voxel spacing: 0.8984 x 0.8984 x 6 mm
- Scanner 3: 5 samples from training and 3 from testing; dimensions: 128 x 128 x 25 pixels; voxel spacing: 1.797 x 1.797 x 5.2 mm
- Scanner 4: 6 samples from training and 1 from testing; dimensions: 192 x 192 x 30 pixels; voxel spacing: 1.25 x 1.25 x 5 mm

Out of the 30 stroke lesions from the training set, 12 were located in the right hemisphere and 18 in the left hemisphere. The size of the stroke varies from small areas affecting only part of a cerebral lobe to larger areas affecting two or even more lobes. They were mainly related to strokes of the anterior circulation, especially in the area supplied by the middle cerebral artery. For that reason, most of the lesions occurred either in the basal ganglia or in the frontal, parietal and temporal lobes.

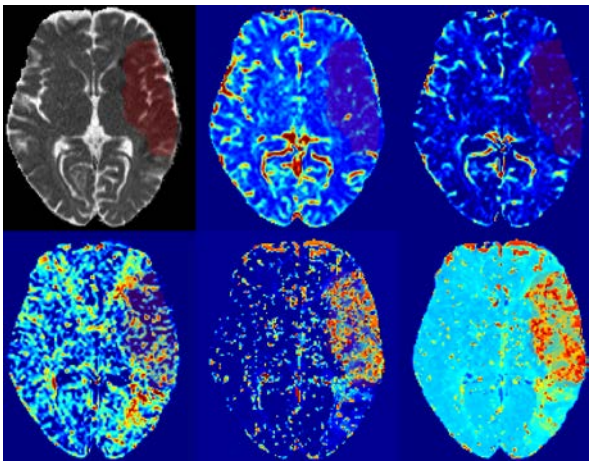


Figure 3. Example of MR sequences of a patient in the training set. Upper part: ADC, CBF and CBV. Lower part: MTT, Tmax and TTP. Ground truth for the lesion segmentation is shown overlayed in red. This particular case shows a large stroke in left hemisphere, affecting frontal and insular lobes and extending to the basal ganglia. The images demonstrate decreased signal in ADC, CBV and CBF with increased signal in MTT, Tmax and TTP in the stroke region.

4.1.2. Clinical Data

The provided clinical data included the following:

- Thrombolysis in Cerebral Infarction (TICI) scale: provides a standardized method to evaluate intracranial perfusion assessed in cerebral angiography (Higashida et al., 2003). It is used to assess the re-perfusion achieved after a flow-restoration intervention such as thrombectomy. Varies from 0 (no restoration of blood flow) to 3 (complete restoration of blood flow). The complete description is shown in Table 2.

Table 2: Description of the TICI scale categories.

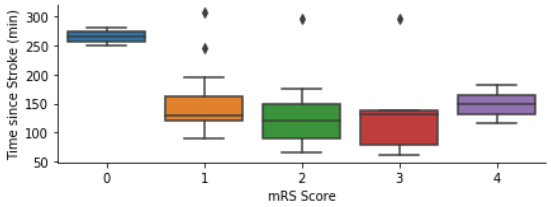
Thrombolysis in cerebral infarction (TICI) Scale	
Category	**Description**
0	No perfusion.
1	Penetration with minimal perfusion.
2	Partial perfusion.
2A	Only partial filling (less than two-thirds) of the entire vascular territory is visualized.
2B	Complete filling of all of the expected vascular territory is visualized but the filling is slower than normal.
3	Complete perfusion.

- Time-since-stroke (TSS): it is expected that the less time passed since stroke onset, the more likely a re-perfusion procedure can salvage brain tissue. In the training set, this feature ranged from 61 to 308 minutes.
- Time-to-treatment (TTT): denotes the time passed between obtaining images and conducting the re-perfusion treatment. Similar to the TSS, it is expected that lower TTTs yield in a higher chance of treatment success. In the training set, this feature ranged from 69 to 221 minutes.
- mRS score: disability score after stroke assessed by the modified Rankin Scale.

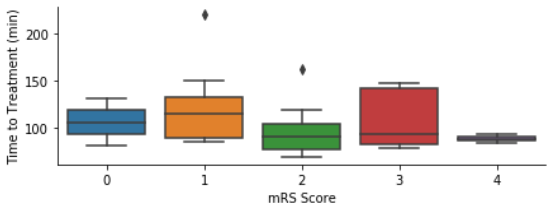
The TICI, TSS and TTT scores were provided for both training and test sets. The mRS score was provided only for the training set, since it was considered as the final outcome for the clinical prediction task. For the training set, there were two patients with score zero (6,67% of the total), thirteen patients with score one (43,3%), eight patients with score two (26,7%), five patients with score three (16,7%) and two patients with score four (6,67%). There were no cases of patients with scores five or six. Distribution of the clinical parameters of the training set and their corresponding mRS scores are summarized in Figure 4.

4.2. *General Pipeline*

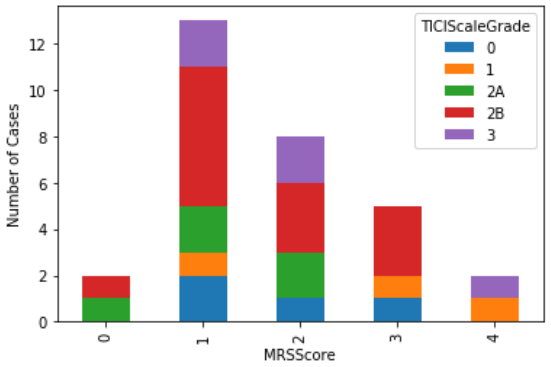
The overall pipeline for mRS classification was composed of three main components: pre-processing, feature extraction and classification. Several pre-processing techniques were used, and not necessarily the same ones were applied in every experiment. Then, features were extracted either using traditional ML approach or a CNN, using either ROIs (segmentation) or the whole brain volume (no segmentation) as input. Finally, three classifiers were trained in different scenarios. The overall framework is depicted in Figure 5. Since it was not possible to assess the mRS scores of the testing set, as the organizers of the Challenge do not accept new submissions anymore, all the experiments were performed using only the training set.



(a) Variation of time since stroke values according to their associated mRS scores.



(b) Variation of time to treatment values according to their associated mRS scores.



(c) Distribution of TICI scale values according to their respective mRS score.

Figure 4. Clinical Data of the ISLES 2016 Challenge training dataset.

4.3. *Pre-processing*

Different preprocessing strategies were used to address particularities of the dataset, including:

- Voxel spacing standardization: as described before, voxel spacing varied between 0.8984 x 0.8984 x 5.0 mm (higher resolution) and 1.797 x 1.797 x 6.5 mm (lower resolution). To make volumes comparable in terms of geometrical features, pixel dimensions were standardized to 1.0 x 1.0 x 6.0 mm through reescaling.
- Removal of empty slices: all cases contained a variable number of slices filled with zeros, which increased memory and computational power needs without adding useful information. Elimination of completely blank slices was performed.
- CBF/CBV difference image creation: for each patient, a 3D difference image was created by subtracting the CBV from the CBF sequences. Those two regions were chosen among all the available ones because they are two of the most significant

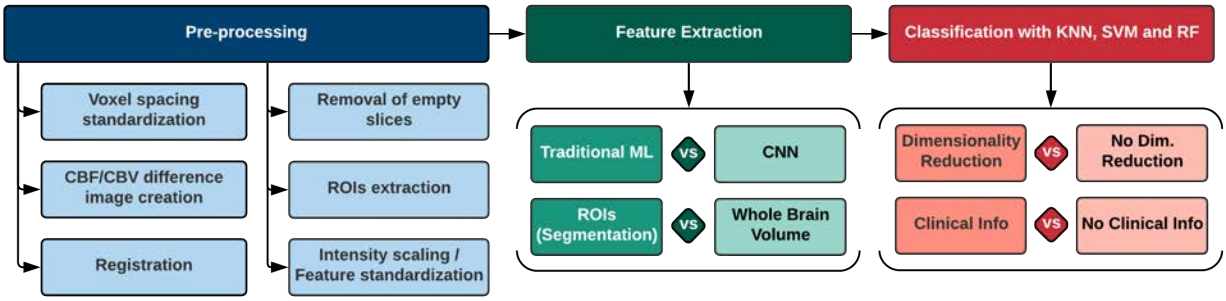


Figure 5. Diagram showing steps implemented in this project.

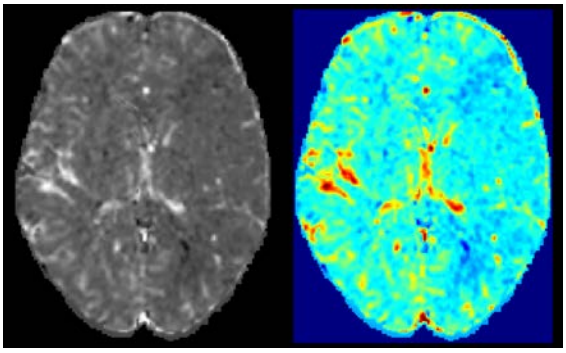


Figure 6. CBF/CBV difference image in two different color maps. Large hypointense area in left hemisphere, corresponding to the stroke region.

to distinguish the core and the penumbra regions of the stroke (Levi et al., 2011), and therefore they carry significant information about the brain's potential to recover from injury. An example of this generated image is shown in Figure 6.

- ROIs extraction: this refers to two types of ROI. For the ML approach, three regions of interest were defined for each patient, following the method described by Maier and Handels (2016), as previously shown in Figure 2: the stroke lesion, the band around the lesion and the rest of the brain. The band around the lesion and the rest of the brain were defined according to a binary mask. For the CNN approach, ROI extraction refers to the creation of a new volume for each patient containing only the images in the binary mask area. The binary mask used to extract the lesion was either the provided ground truth (to have a baseline of the optimal results achievable by a perfect segmentation) or a segmentation mask provided by Clérigues et al. (2018), who implemented deep learning strategies for automated stroke lesion outcome segmentation. The authors used a patch based strategy to create probability maps of voxels belonging to either lesion or healthy tissue using a 3D U-net. For each case, two segmentation masks (referred to as Segmentation 01 and Segmentation 02 in this project) were generated, by varying the threshold in the probability map to define if a voxel corresponds to lesion or to normal brain tissue. Segmentation 01 was generated using a higher threshold and achieved a higher DICE similarity score (38.8%), but sometimes created small stroke areas. Segmentation 02, on the other hand, used a lower threshold, resulting in larger stroke areas, although with poorer performance in terms of DICE score (36.5%).
- Registration: although the patients were co-registered (all sequences for one patient were in the same space), the provided dataset had no inter-patient registration. Two cases were used as reference space for the registration of the other patients: cases 15 and 19. Those two cases were chosen because they cover most of the brain area, extending from the posterior fossa to the high convexity, and also because they were better aligned to the midline. This was useful to extract features related to the symmetry between hemispheres.
- Intensity scaling: since the exams were acquired in different scanners, there were differences in terms of intensities for each sequence. Intensity scaling from 0 to 1 for each sequence was performed to tackle this issue.
- Feature standardization: as a final preprocessing step, feature standardization was performed by removing the mean and scaling to unit variance. The statistics used to guide this process were obtained only from the training samples.

4.4. Traditional Machine Learning Approach

The traditional ML approach is based on extracting meaningful hand crafted imaging features that are used to perform classification in target groups. For this project, different sets of features were extracted, as described below. The ML strategies are summarized in Figure 7.

4.4.1. Feature Extraction from ROIs

The initial approach of this project was to extract features from ROIs in a similar way as described by Maier and Handels (2016), that represents the current state of the art in this challenge. From the ADC images of those

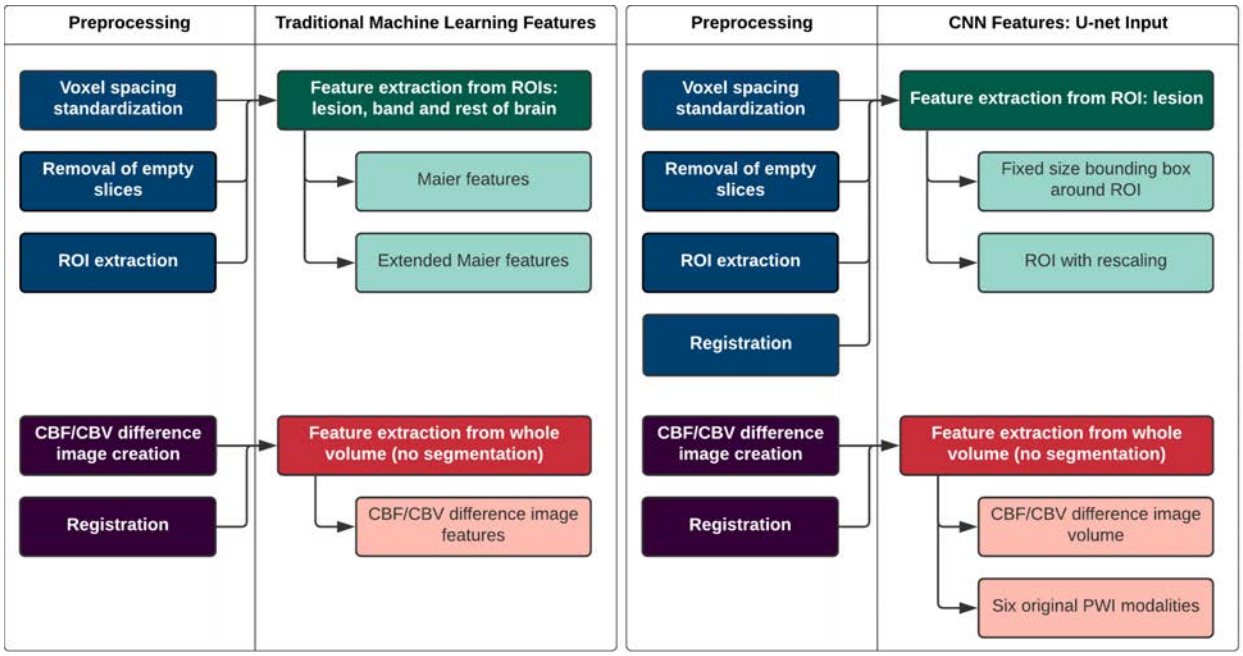


Figure 7. Overview of different strategies used for the traditional machine learning approach (left) and the convolutional neural network approach (right).

regions, the same features described by Maier and Handels (2016) were extracted. Additionally, considering that the CBV and the CBF are two of the most significant sequences in MR perfusion, another set of experiments was performed by extracting the same set features, not only from the ADC, but also extracting intensity features from those two sequences (referred as Extended Maier Features in this master thesis).

4.4.2. Feature Extraction from the Whole Brain Volume

A different approach was developed without requiring lesion segmentation. From each slice of the CBF/CBV difference image generated during the preprocessing stage, the absolute subtraction between left and right hemisphere was obtained. This highlights the difference in perfusion between left and right hemispheres. It is expected that a healthy brain shows similar blood circulation in both hemispheres, whereas a patient with a large stroke should show one hemisphere with lower perfusion in comparison to the opposite one. Also, in the affected brain area, the difference between those two sequences may carry information about the potential recovery of the tissue in that voxel, since the core usually shows a decrease in CBV and CBF and the penumbra shows a decrease in CBF with relative preservation of the CBV.

Intensity features similar to the ones implemented by Maier and Handels (2016) were then extracted from this newly obtained image and used for classification. Additionally, the same process was applied to a Gaussian smoothed version of the difference image, with two different values for sigma (one or three), to reduce the effect of misalignment between voxels when performing subtraction between left and right hemispheres. This process is shown in Figure 8.

4.5. Convolutional Neural Networks

4.5.1. U-Net and 3D Unet

The U-Net architecture was first proposed and implemented by Ronneberger et al. (2015) for the seg-

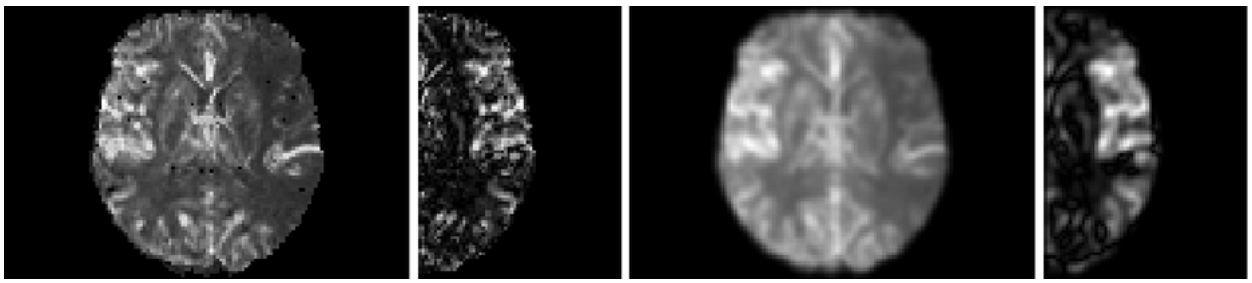


Figure 8. Whole Image Features: absolute difference of brain hemispheres from the CBF/CBV difference image and from the Gaussian smoothed version of the same image.

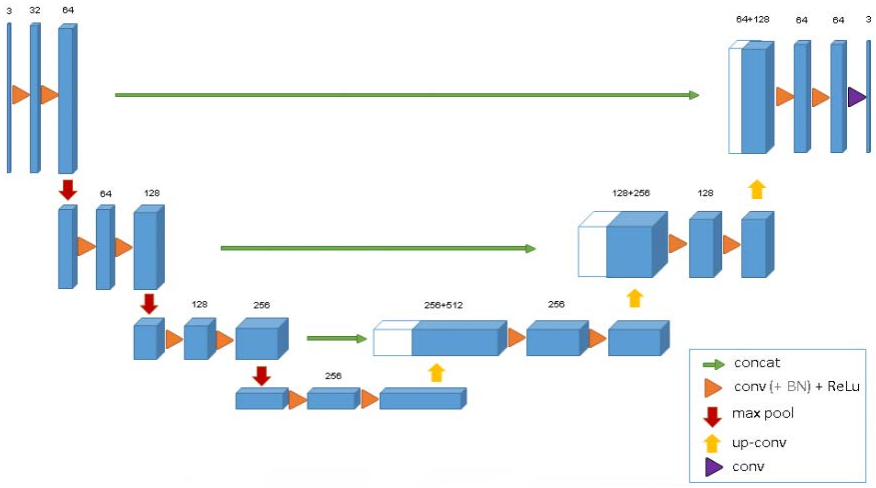


Figure 9. 3D U-net architecture (Çiçek et al., 2016).

mentation of neuronal structures. Using that architecture, the authors won the cell tracking Challenge proposed during the International Symposium on Biomedical Imaging (ISBI) 2015. It consists of three parts: (1) encoder/downsampling; (2) bottleneck; (3) decoder/upsampling. It has a symmetrical shape, resembling the letter 'U', which explains its name. The encoding part is composed of four blocks, where each block is formed by two 3 x 3 convolutional layers (each followed by a ReLU activation function) and by a 2 x 2 max pooling layer with stride 2. The bottleneck happens between the contracting and expanding paths, and it is built with two convolutional layers and a dropout layer. Finally, the decoding part is also composed of four blocks, each of them including a deconvolution layer with a stride 2, a concatenation step with the corresponding cropped feature map of the contracting path and two 3 x 3 convolutional layers (with ReLU). This architecture has proven to achieve great performance in the segmentation task of different types of biomedical images, such as intervertebral disks segmentation (Liu, 2018) and brain segmentation (Luna and Park, 2018) for 2018 MICCAI Challenges.

The 3D U-net is a modification of the original U-net architecture that was proposed by Çiçek et al. (2016) to be used for volumetric segmentation that learns from sparsely annotated volumetric images. The network proposed by this author follows the same structure as the one described above, with the difference that it takes 3D volumes as input and processes them with 3D operations (3D convolutions, 3D max pooling and 3D up-convolutional layers). Other differences include removal of one resolution step and the addition of a batch normalization layer at each step. A visual representation of the 3D U-net is shown in Figure 9.

4.5.2. Feature Extraction using a 3D U-net

Since this project tackles a classification problem rather than a segmentation one, it was hypothesised that the 3D U-net could still extract meaningful image characteristics to achieve the final objective. The goal of this approach was to use the encoder part of the architecture to extract high level features. The ouput of the blocks of convolutional layers was flattened to be used as a feature vector for later classification. The network was pretrained for the segmentation task of the images of this same dataset (Clérigues et al., 2018), and the weights were used to perform the feature extraction. The analyzed strategies are summarized in Figure 7.

For the experiments using ROIs, a bounding box was defined around the segmentation mask, and a new volume for each patient was created, where each channel included information about a sequence (ADC, CBV, CBF, MTT, Tmax, TTP) only inside the limits defined by that box. This new volume was then used as the network input.

For the other experiments, the whole brain volume was used a the network input. This was either a volume containing six channels (one for each sequence: ADC, CBV, CBF, MTT, Tmax, TTP) or a volume containing only one channel (experiments using CBF/CBV difference image).

4.6. Classification

Three different classifiers were analyzed for every experiment: K-nearest neighbor (KNN), support vector machines (SVM) and random forests (RF). An ensemble strategy was also implemented. The effect of introducing dimensionality reduction and clinical information was also analyzed.

Principal component analysis (PCA) was used to perform linear dimensionality reduction using singular value decomposition of the data in the feature vector to

project it to a lower dimensional space, therefore reducing the length of the feature vector.

Clinical information provided for the training cases was incorporated as additional entries for the corresponding feature vector of each patient.

4.6.1. Classifiers

The KNN algorithm is based on the assumption that similar examples exist in close proximity in the feature space. That way, when a new sample is introduced to the system, the classifier checks what are the labels of the nearest data points and assigns the label that occurs the most to this new sample. Significant hyperparameters include number of neighbors and class weights. It is one of the most simple classifiers in machine learning, and can be considered a naive classifier.

SVM consist of a learning model for classification and regression problems. Given a set of features and their corresponding classes, this classifier builds a hyperplane to separate features related to different classes. The best result is given by the maximum margin hyperplane, named the optimal separating hyperplane (OSH). The linear discriminant function of this OSH is called support vector machine. Important hyperparameters in this model include kernels types (linear, rbf, sigmoid, polynomial), penalty for errors (values of C) and class weights.

RF classifier creates an ensemble of independent decision trees, where every tree is built by random selection of features (Breiman, 2001), each tree being a weak classifier. The final decision is obtained as a result of majority voting. The hyperparameters to be tuned include max depth of the forest, number of trees and class weights.

Ensemble strategies combine the output of different classifiers to improve generalization when compared to a single estimator. In this project, the outcomes of the best obtained models were combined using majority voting.

4.7. Validation and Statistical Analysis

The hyperparameters of the classifiers were tuned through grid searches to minimize the MAE between the predicted and the actual mRS scores for each case, which was the main goal of this project. Additionally, the models were also evaluated through their confusion matrices.

The cross-validation strategy used was leave-one-out (Cawley and Talbot, 2003). In cross-validation, the available data is divided into k disjoint sets, where k models are then trained, each on a different combination of $k - 1$ partitions and tested on the remaining partition. Leave-one-out is the extreme case where k is equal to the number of samples in dataset.

Statistical analysis in groups with unequal sample size was performed using Mann–Whitney U test. Paired groups were compared using Wilcoxon signed-rank test. A p value lower than 0.05 was considered significant.

4.8. Implementation Details

This project was implemented using Python programming language. Supportive libraries used include numpy and matplotlib. Extraction of hand crafted features was done using scikit-image package (Van der Walt et al., 2014). Completion of missing clinical information entries was performed using fancyimpute package (Rubinsteyn and Feldman). The 3D U-net was implemented using PyTorch (Pfeiffer, 2007). The Scikit-learn library (Klikauer, 2016) was used to perform the classification task and to obtain evaluation metrics. Data analysis was done using pandas (McKinney, 2011) and SciPy (Jones et al., 2001) libraries. ITK-SNAP (Yushkevich et al., 2006) was used for image visualiztion. Image registration was performed with elastix (Klein et al., 2010). The project is available at https://github.com/marcioabcr/ClinicalPredict.

4.9. Specific Objectives

The quantitative performance of the different models was evaluated using the training set of the Challenge through assessment of the mean and standard deviation of the absolute error of the mRS score. The main objectives of the initial experiments were:

- Evaluate the performance of models using different classifiers.
- Evaluate the performance of models regarding usage of dimensionality reduction through PCA and usage of clinical information.

According to these initial results, further analysis of the implemented strategies was performed, this time keeping only the experiments with the two classifiers with the best performance and only keeping the experiments with the parameters that obtained the best scores in terms of dimensionality reduction and clinical information usage (with or without PCA and with or without clinical information incorporated).

By keeping only the classifiers and the parameters with the best results, this second round of analysis focused on comparing the performance between main approaches (traditional ML vs. CNN and ROI-based vs. whole brain volume) to deal with the project objective. More specifically, the objectives were to:

- Evaluate the performance of ROI-based strategies regarding usage of the ground truth and usage of a mask provided by an automated segmentation method.
- Evaluate the best results obtained for each of the categories previously described (traditional ML and CNN approaches, either using ROIs or whole brain volume as input) in terms of mean and standard deviation of the absolute error.

Table 3: Means and standard deviations (Std) of absolute errors obtained for different strategies using segmentation (ML - traditional machine learning; CNN - convolutional neural network; RF - random forests; SVM - support vector machines; ROI - region of interest).

Absolute Errors			Classifier	
			RF	SVM
Approach	Features	Segmentation	Mean ± Std	Mean ± Std
ML	Maier	Ground truth	0.40 ± 0.66	0.43 ± 0.72
		Segmentation 01	0.53 ± 0.72	0.60 ± 0.99
		Segmentation 02	0.63 ± 0.71	0.60 ± 0.99
	Extended Maier	Ground truth	0.43 ± 0.67	0.57 ± 0.67
		Segmentation 01	0.60 ± 0.71	0.60 ± 0.99
		Segmentation 02	0.67 ± 0.75	0.60 ± 0.99
CNN	3D U-net	Ground truth	0.77 ± 0.88	0.58 ± 0.99
		Segmentation 01	0.73 ± 0.81	0.59 ± 0.99
		Segmentation 02	0.78 ± 0.80	0.60 ± 0.99

- Evaluate the confusion matrix and the accuracy of the models that achieved best performance for traditional ML and CNN approaches.
- Evaluate the performance of the ensemble approach

5. Results

5.1. Overview

The experiments were divided in five groups, according to the strategy employed. First three groups were related to traditional ML approaches: (1) feature extraction from ROIs - Maier features; (2) feature extraction from ROIs - Extended Maier Features (Maier features applied to ADC, CBV and CBF); (3) feature extraction from the whole brain volume (CBF/CBV difference image features). The last two groups were related to CNN-based approaches: (4) feature extraction from ROI; (5) feature extraction from whole brain volume.

5.2. Comparison between model configurations

5.2.1. Naive Classifier vs. SVM/RF

Experiments using KNN as the classifier had a mean absolute error of 0.92 ± 0.88, while RF and SVM obtained a value of 0.72 ± 0.82 and 0.57 ± 0.96, respectively. This difference in performance is statistically significant both comparing KNN with RF and KNN with SVM ($p < 0.05$).

If we analyze separately traditional ML and CNN strategies, this statistically significant difference also holds. For ML experiments, the MAE was 0.9 ± 0.88 for KNN, 0.65 ± 0.78 for RF and 0.55 ± 0.93 for SVM. For CNN experiments, the MAE was 0.93 ± 0.88 for KNN, 0.76 ± 0.85 for RF and 0.58 ± 0.99 for SVM.

5.2.2. Dimensionality Reduction and Usage of Clinical Information

Experiments using dimensionality reduction of the feature vector through PCA showed a MAE of 0.66 ± 0.90, while experiments without it had a MAE of 0.63 ± 0.90. If we consider only CNN experiments, the MAE was respectively 0.68 ± 0.92 with PCA and 0.67 ± 0.93 without it. None of these differences was statistically significant. For ML experiments, the MAE was 0.63 ± 0.87 with dimensionality reduction and 0.57±0.84 without ($p < 0.05$).

Experiments adding clinical information to the feature vector had a MAE of 0.64 ± 0.90, while experiments without it had a MAE of 0.65 ± 0.90. In ML cases, the MAE was 0.60 ± 0.86 using clinical information and 0.61 ± 0.86 not using it. In CNN experiments, the MAE was 0.67 ± 0.93 using clinical information and 0.67 ± 0.92 without it. None of those differences were statistically significant.

5.3. Comparison between approaches

For further analysis, experiments that used KNN classifier were discarded. Similarly, only experiments without dimensionality reduction and with clinical information were kept, since experiments without dimensionality reduction had a better or equivalent performance in comparison with the cases using it and the experiments with or without clinical information had comparable performance.

5.3.1. Ground truth vs. Automated Segmentation

The results of segmentation strategies are summarized in Table 3. In the cases where the original Maier features were extracted, the usage of ground truth masks resulted in a MAE of 0.40 ± 0.66 using RF

Table 4: Mean and standard deviations (Std) of absolute errors obtained for different strategies that achieved the best results (ML - traditional machine learning; CNN - convolutional neural network; RF - random forests; SVM - support vector machines; ROI - region of interest).

Absolute Errors			Classifier	
			RF	SVM
Approach	Segmentation	Features	Mean ± Std	Mean ± Std
ML	Yes	Maier	0.53 ± 0.72	0.60 ± 0.99
		Extended Maier	0.60 ± 0.71	0.60 ± 0.99
	No	CBF/CBV Diff Image	0.47 ± 0.67	0.43 ± 0.76
CNN	Yes	ROI	0.73 ± 0.81	0.57 ± 0.99
	No	Whole volume	0.73 ± 0.81	0.50 ± 0.96

and 0.43 ± 0.72 using SVM, whereas Segmentation 01 achieved 0.53 ± 0.72 using RF and 0.60 ± 0.99 using SVM and Segmentation 02 reached 0.63 ± 0.71 with RF and 0.60 ± 0.99 with SVM. The cases that employed Extended Maier features achieved similar or worse performance in comparison with the original Maier characteristics.

For approaches using feature extraction with CNNs, the scores were similar for both types of segmentation and for the ground truth, with experiments using SVM obtaining a lower MAE. The only comparison that showed a statistically significant difference was between ground truth and Segmentation 02 using original Maier features and RF as a classifier, where the ground truth had a superior performance ($p < 0.05$).

5.3.2. *Best Individual Results*

The best results obtained for each approach and for each feature extraction method are shown in Table 4. The best results for the traditional ML strategy were obtained extracting hand crafted features from the CBF/CBV difference image, with registration to the case 15 as a preprocessing step, applying a Gaussian smoothing with sigma equals to three and using SVM as a classifier. This method achieved a MAE of 0.43 ± 0.76 and an accuracy of 70%. The corresponding confusion matrix is shown in Figure 10a, displaying the distribution of the predicted scores and the actual labels.

For the CNN strategy, the best results were obtained extracting whole volume features also from the CBV/CBF difference image, but this time registered to the case number 19. The classifier was also SVM. This strategy resulted in a MAE of 0.50 ± 0.96 and in an accuracy of 76.67%. The associated confusion matrix is shown in Figure 10b.

5.3.3. *Classifier Ensemble*

From the strategies displayed in Table 4, an ensemble method was built, combining the outcome of two ML models (the one using Maier features and the one using CBF/CBV difference image features) and the two CNN methods. Using majority voting, the ensemble method achieved a mean absolute error of 0.60 ± 0.99 and 70% accuracy. The associated confusion matrix is shown in Figure 10c.

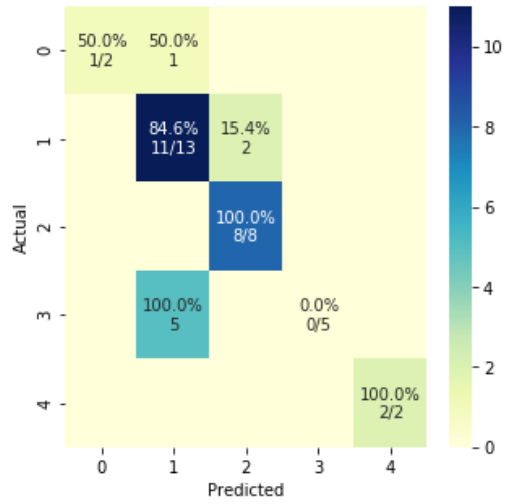
6. Discussion

Throughout this work, it was possible to fulfil the initial goal of implementing strategies to predict clinical outcome from MR images and from clinical data. Also, given the large number of performed experiments and the different pipelines used, it is possible to make some assumptions about the achieved results.

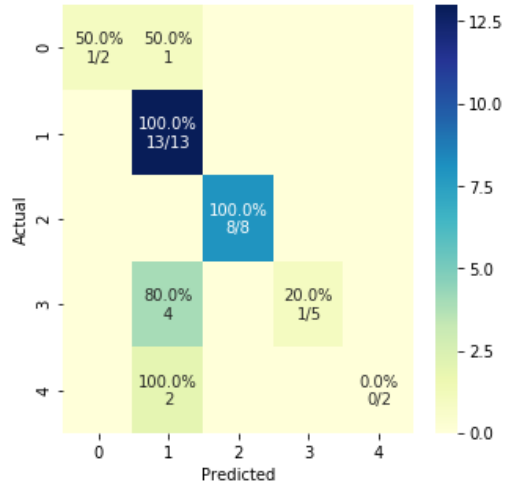
For the types of classifiers that were trained, KNN had the worst performance, with a mean absolute error close to one and a large standard deviation. This was expected, since it is one the most simple classifiers in machine learning, unsuitable for such a complex task.

In terms of performance between random forests and support vector machines, it is noticeable that they achieved similar results regarding traditional ML approach, with RF obtaining better results using Maier features and SVM using features from the CBF/CBV difference image. For the CNN approach, the performance of SVM was superior to RF. This difference can be explained by the nature of each classifier. Random forests use a random selection of features to build the decision trees. The approaches using CNNs present a significantly larger feature vector, making it much harder to select significant features to build the classifier.

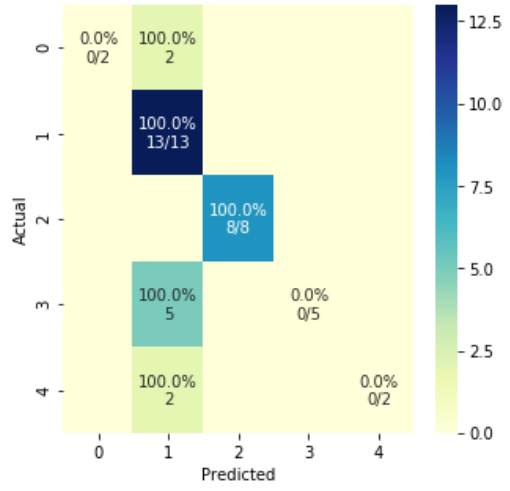
Experiments using dimensionality reduction showed either comparable or worst performance compared to experiments without it. This might be due to the fact that, in scikit-learn, the number of components chosen to keep when applying PCA is limited by the number of samples. Since the dataset contained only 30 samples, the size of the feature vector after using PCA was limited to 29 (since one case was kept for testing in leave-



(a) Confusion matrix of the best result using traditional ML strategy.



(b) Confusion matrix of the best result using CNN strategy.



(c) Confusion matrix of ensemble strategy.

Figure 10. Confusion matrices of different approaches

one-out strategy). This feature vector size is probably insufficient to encode the complexity of this task.

Regarding usage of clinical information, it did not provide any significant change in the model's performance. By analyzing Figure 4, it is evident that the initial assumptions about how the provided clinical parameters might affect stroke outcome do not seem to be true. For example, in Figure 4a it is possible to observe that patients with mRS score zero had time since stroke values much higher than cases with mRS score of four. The same mismatch occurs for time to treatment parameter. This is another evidence that shows how it is not so straight forward to predict clinical outcome based only in a handful of data. Probably the addition of more relevant clinical parameters such as gender, age, body mass index (BMI) and other risk factors for cardiovascular diseases would be benefitial. As previously shown by Kabir et al. (2017) and Vrtkova (2017), the use of clinical information alone gathered from the EHR already showed promising results to predict clinical outcome. Merging more meaningful clinical parameters with imaging data could improve the results even further.

There are several limitations regarding the CNN approach for this task. First of all, the network was not trained with the purpose of clinical outcome prediction. There was no available network that was pretrained for this purpose and with the specific sequences provided in the dataset, preventing us to directly apply transfer learning. Since clinical outcome depends on the interaction of several factors, we hypothesized that we would not be able to use a patch based approach to train the network, since it is important to obtain information from the entire volume to be able to achieve this complex goal. This issue could be minimized if a large dataset was available, but since only 30 cases were provided, training a network from scratch was not possible. To try to overcome this problem, it was decided to use weights that come from a network trained with the same images, but with a different task (patch based segmentation). Extracted features using this method may not be able to depict the most meaningful information for the task at hand.

Traditional ML strategies achieved a lower mean absolute error when compared to the CNN approach. While it is expected that deep learning has the ability to surpass the vast majority of machine learning approaches when dealing with a large number of samples, in this scenario, with only 30 samples, the use of hand crafted features is still very useful and achieves a significantly good performance. While the ideal size of the feature vector depends on the type of the data and the classifier used (Hua et al., 2005), generally a feature vector that outnumbers the sample size leads to a poor performance. In this project, all strategies had a feature vector larger than the sample size, since the number of cases was only 30. However, the CNN approaches re-

sulted in a much larger feature vector than the machine learning ones, which might also depreciate the performance.

Strategies that used lesion segmentation and strategies without it achieved similar performance. In fact, the overall best results were obtained without use of segmentation, with an error of 0.43 ± 0.76 in the machine learning approach and 0.50 ± 0.96 for the CNN approach. It is interesting to develop such strategies, since this makes sure that eventual errors in segmentation are not propagated to the clinical outcome algorithm. Furthermore, it is possible to eliminate an intermediate step, which adds robustness to the process, while also making it faster. This also eliminates concerns such as how to proceed if the segmentation algorithm fails or if it does not provide a valid mask. When using a strategy based on segmentation, Table 3 shows how the quality of the segmentation is essential for the model performance. When using ground truth as a mask, some of the best overall results were obtained, specially in the traditional machine learning strategy. When using Segmentation 01, the performance already drops. Finally, Segmentation 02 has a significantly lower performance when compared to the ideal scenario (ground truth).

The strategy without lesion segmentation uses a newly generated image from the difference between CBF and CBV. This derives from prior knowledge about the research topic (stroke imaging), which highlights how the use of radiological information can be important to guide what is the most relevant information that can be taken from the dataset in order to achieve the predefined goal. If the dataset was large enough, deep neural networks could overcome this issue, and be able to independently select the appropriate features from the provided images.

Looking at the best obtained results, it is possible to observe that the confusion matrices shown in Figures 10a and 10b show similar performance in terms of mean absolute error, but differences on the error pattern. While the CNN approach missed all category four cases (0% accuracy for this class) and correctly assigned one out of the five cases for category three (20% accuracy), the machine learning correctly classified both category four cases (100% accuracy). The overall error was similar because the machine learning strategy made small mistakes in terms of assigning label two to class one (84.6% accuracy) and because it did not get any of the category three cases correct (0% accuracy). It is important to discuss which of those errors should be more strongly avoided in a real clinical scenario. Probably it would be more adequate to allow small mistakes in categories that represent a better clinical outcome than to misclassify cases with a more negative result. This also demonstrates why in such tasks the accuracy is not the only metric to be observed, since the class imbalance would bias the classification to the more prevalent categories, which could lead to a disaster in terms of patient management.

The results of the classifier ensemble strategy is a good example of this issue. While achieving a small mean absolute error of 0.60 ± 0.99 and achieving perfect prediction for patients in categories one and two (the most prevalent ones), it failed to classify all the other categories (zero, three and four). In fact, it predicted label one for all the other cases, as we can see in Figure 10c. This is an undesirable case, where, due to the small dataset and the large data imbalance, the classifier just decides to label every new case to the most prevalent class.

One possible strategy to tackle this issue is the use of a different evaluation metric to guide the tuning of the classifier's hyperparameters and to assess the overall performance. This project was based on the ISLES Challenge 2016, and therefore it kept the metric that was established by it. Instead of the mean absolute error, using root mean squared error, for example, could minimize the problem mentioned above. By taking into account also the magnitude of the error, this could help prevent the classifier to just assign the label of the most prevalent class. Another significant improvement could be achieved if the dataset included more cases from the less prevalent scores, diminishing the class imbalance.

In terms of comparison to the state of the art, it is difficult to make an assertion, since Maier and Handels (2016) described only the results they obtained in the 19 images of the testing set, not in the training set. In the end of this project, an attempt was made to send the results obtained by applying our best model to the testing set, but the organizers of the Challenge do not provide support to new submissions anymore.

It is important to note the significant limitations of the provided ISLES dataset. The task of predicting clinical outcome is one of the hardest in the medical domain, and depends on a combination of several factors. That means that it is necessary to gather a large amount of good quality data in order to build a robust model that can be actually used in clinical practice.

In terms of quantity, the dataset contains only 30 patients in a classification problem with five classes. Also, the distribution of patients is imbalanced. The majority of samples (21 cases, or 70%) belong to two labels (mRS scores one and two), while for two labels (mRS scores zero and four), there are only two samples (or 6,7%) each. For labels five and six, cases were not provided. As previously stated, this class imbalance introduces a bias when building classifiers, making it harder for them to learn what are the characteristics of the classes with a lower number of samples. Also, due to the nature of the task, it is difficult to think in terms of data augmentation, since the clinical outcome does not depend only on image characteristics of a single modality or only on clinical information, but it depends on the combination of all those types of information.

In terms of quality, there are some aspects that can be

discussed. The dataset does not include DWI sequences, which is one of the most important ones for stroke evaluation. This sequence has extremely high sensitivity for the diagnosis of stroke, showing changes in an earlier stage compared to other modalities and other sequences. In clinical practice, DWI and ADC are always obtained when a stroke is suspected, because both images are necessary for the correct diagnosis. Also, additional images such as NCCT and angiotomography of brain vessels could potentially increase the performance of the model.

The fact that the images were acquired from different scanners adds another challenge to the task. Since each machine leads to differences in intensity, dimension and voxel spacing, it is necessary to tackle this issue. This inevitably leads to interpolations, change in resolution and loss of possibly relevant information.

On top of all the previously discussed dataset characteristics, another limitation that should be addressed is that all cases are related to ischemic strokes of the anterior circulation. Even though this type is the most prevalent, the addition of cases from posterior circulation would be fundamental to make the model more robust, with a higher generalization capability and more suitable for a future use in clinical practice.

7. Conclusions

In this project, we proposed different strategies to predict clinical outcome of patients with acute ischemic stroke. This work was motivated by the ISLES Challenge 2016, which provided the dataset containing MR images and clinical information. The approaches varied in terms of preprocessing steps, feature extraction methods (traditional ML versus CNN-based), inputs to the model (ROIs versus whole brain volume) and classification parameters (types of classifiers, dimensionality reduction techniques and clinical information usage). The final goal was to minimize the mean absolute error when evaluating the modified Rankin Scale score, which assesses patient's degree of disability after a stroke event by assigning a score from zero to six.

In terms of classifiers, KNN had a significantly worse performance than SVM and RF. Strategies using dimensionality reduction with PCA achieved similar or worse results than strategies that did not use it, and there was no statistically significant difference between adding clinical information or not to the feature vector.

For the traditional ML approach, the set of hand crafted features that was used in experiments with ROIs was inspired by Maier and Handels (2016), which represents the current state of the art for this task, while the features used in whole brain volume approaches was obtained from a 3D volume generated from the subtraction of the CBV from the CBF perfusion sequences. The best result obtained from the first strategy was a mean absolute error of 0.53 ± 0.72 (when using a binary mask provided by an automated lesion segmentation method (Clérigues et al., 2018)), while the second strategy achieved 0.43 ± 0.76.

For the CNN approach, feature extraction was performed using the encoder part of a 3D U-net, which was pretrained to perform patch-based lesion segmentation of stroke on the same images (Clérigues et al., 2018). When using ROIs as network input, the best obtained result was a mean absolute error of 0.57 ± 0.99, and the best score for the whole volume strategy was 0.50 ± 0.96 when using the CBF/CBV difference volume as input.

An ensemble strategy combining the best models through majority voting was also implemented, achieving a mean absolute error of 0.60 ± 0.99. This approach, however, correctly classified only the samples belonging to the most prevalent groups (scores one and two), while it failed to assign the correct label to the least prevalent groups.

The small number of samples included in the dataset (only 30 patients) constitutes a significant limitation to the proper evaluation of the implemented strategies. However, it is possible to highlight how traditional machine learning strategies still achieve good performance when dealing with a limited dataset, which is usually the case in the biomedical setting. Also, it is important to notice that strategies that did not require lesion segmentation achieved similar or even better results than strategies that used it. This shows how radiological knowledge can be incorporated into project design, in order to guide the extraction of the most meaningful information from the available resources.

This work is a first step in dealing with clinical outcome prediction, which is one of the next frontiers to be reached by new machine learning and deep learning algorithms in the medical scenario. The initial results are promising, and future work with larger datasets is imperative to make this type of algorithm usable in a real clinical setting, ultimately providing useful information to better guide patient management.

8. Acknowledgments

We thank Albert Clérigues (University of Girona) for providing lesion segmentation masks and technical support for the impletation of the 3D U-net. We thank Sergi Valverde and Kaisar Kushibar (University of Girona) for assisting in project design and implementation. We thank Maria Gabriela Longo (Radiology Department, Massachusetts General Hospital) for assisting in the statistical analysis. We thank Mladen Rakic (University of Girona) for helping in reviewing the manuscript. Finally, we thank VICOROB (Computer Vision and Robotics Group, University of Girona) for providing the resources needed for the implementation of this project.

References

Allmendinger, A.M., Tang, E.R., Lui, Y.W., Spektor, V., 2012. Imaging of stroke: Part 1, Perfusion CT–overview of imaging technique, interpretation pearls, and common pitfalls. American Journal of Roentgenology 198, 52–62.

Bacchi, S., Zerner, T., Oakden-Rayner, L., Kleinig, T., Patel, S., Jannes, J., 2019. Deep Learning in the Prediction of Ischaemic Stroke Thrombolysis Functional Outcomes: A Pilot Study. Academic Radiology .

Barow, E., Boutitie, F., Cheng, B., Cho, T.H., Ebinger, M., Endres, M., Fiebach, J.B., Fiehler, J., Ford, I., Galinovic, I., Nickel, A., Puig, J., Roy, P., Wouters, A., Thijs, V., Lemmens, R., Muir, K.W., Nighoghossian, N., Pedraza, S., Simonsen, C.Z., Gerloff, C., Thomalla, G., 2019. Clinical Characteristics and Outcome of Patients with Lacunar Infarcts and Concurrent Embolic Ischemic Lesions. Clinical Neuroradiology .

Benjamin, E.J., Virani, S.S., Callaway, C.W., Chamberlain, A.M., Chang, A.R., Cheng, S., Chiuve, S.E., Cushman, M., Delling, F.N., Deo, R., De Ferranti, S.D., Ferguson, J.F., Fornage, M., Gillespie, C., Isasi, C.R., Jiménez, M.C., Jordan, L.C., Judd, S.E., Lackland, D., Lichtman, J.H., Lisabeth, L., Liu, S., Longenecker, C.T., Lutsey, P.L., MacKey, J.S., Matchar, D.B., Matsushita, K., Mussolino, M.E., Nasir, K., O'Flaherty, M., Palaniappan, L.P., Pandey, A., Pandey, D.K., Reeves, M.J., Ritchey, M.D., Rodriguez, C.J., Roth, G.A., Rosamond, W.D., Sampson, U.K., Satou, G.M., Shah, S.H., Spartano, N.L., Tirschwell, D.L., Tsao, C.W., Voeks, J.H., Willey, J.Z., Wilkins, J.T., Wu, J.H., Alger, H.M., Wong, S.S., Muntner, P., 2018. Heart disease and stroke statistics - 2018 update: A report from the American Heart Association. volume 137. doi:10.1161/CIR.0000000000000558, arXiv:NIHMS150003.

Bone, I., Hareendran, A., Grant, M., Muir, K.W., Wilson, J.L., Baird, T., Schulz, U.G., 2002. Improving the Assessment of Outcomes in Stroke. Stroke 33, 2243–2246.

Brea, A., Laclaustra, M., Martorell, E., Pedragosa, À., 2013. Epidemiología de la enfermedad vascular cerebral en España. Clinica e Investigacion en Arteriosclerosis 25, 211–217.

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32. URL: https://doi.org/10.1023/A:1010933404324, doi:10.1023/A:1010933404324.

Calamante, F., Christensen, S., Desmond, P.M., Ostergaard, L., Davis, S.M., Connelly, A., 2010. The physiological significance of the time-to-maximum (Tmax) parameter in perfusion MRI. Stroke 41, 1169–1174.

Cawley, G.C., Talbot, N.L., 2003. Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. Pattern Recognition 36, 2585–2592.

Choi, Y., Kwon, Y., Lee, H., Paik, M.C., Won, J.H., 2016. Deep Convolutional Neural network Approach for Brain Lesion Segmentation. Ischemic Stroke Lesion Segmentation (ISLES) 2016 URL: http://www.isles-challenge.org/ISLES2016.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-net: Learning dense volumetric segmentation from sparse annotation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9901 LNCS, 424–432. doi:10.1007/978-3-319-46723-8_49, arXiv:1606.06650.

Clérigues, A., Valverde, S., Oliver, A., Lladó, X., 2018. Deep learning architectures for stroke lesion segmentation and outcome prediction, in: Master Thesis Proceedings - International Programme in Vision and Robotics (VIBOT) Day 2018. URL: http://vibotudg.weebly.com/vibot-day-2018.html.

Edlow Jonathan, A., Warach, S., Jaigobin, C., Wardlaw, J., Sandercock, P., Schellinger, P., Brazzelli, M.G., 2011. Evidence-based guideline: The role of diffusion and perfusion MRI for the diagnosis of acute ischemic stroke: Report of the Therapeutics and Technology Subcommittee of the American Academy of Neurology. Neurology 76, 2036–2038. doi:10.1212/wnl.0b013e318219a0b4.

Egger, K., Maier, O., Reyes, M., Wiest, R., 2016. Ischemic Stroke Lesion Segmentation (ISLES) Challenge 2016. URL: http://www.isles-challenge.org/ISLES2016/.

Favate, A.S., Younger, D.S., 2016. Epidemiology of Ischemic Stroke. Neurologic Clinics 34, 967–980. URL: http://dx.doi.org/10.1016/j.ncl.2016.06.013.

Hankey, G.J., 2017. Stroke. The Lancet 389, 641–654.

Hankey, G.J., Blacker, D.J., 2015. Is it a stroke? The BMJ (Online) 350, 1–6.

Higashida, R.T., Furlan, A.J., Roberts, H., Tomsick, T., Connors, B., Barr, J., Dillon, W., Warach, S., Broderick, J., Tilley, B., Sacks, D., 2003. Trial Design and Reporting Standards for Intraarterial Cerebral Thrombolysis for Acute Ischemic Stroke. Stroke 34, e109–e137.

Hua, J., Xiong, Z., Lowey, J., Suh, E., Dougherty, E.R., 2005. Optimal number of features as a function of sample size for various classification rules. Bioinformatics 21, 1509–1515.

Jones, E., Oliphant, T., Peterson, P., 2001. SciPy: Open source scientific tools for Python.

Kabir, A., Ruiz, C., Alvarez, S.A., Moonis, M., 2017. Predicting Outcome of Ischemic Stroke Patients using Bootstrap Aggregating with M5 Model Trees. Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2017) , 178–187.

Klein, S., Staring, M., Murphy, K., Viergever, M., Pluim, J., 2010. elastix: A Toolbox for Intensity-Based Medical Image Registration. Institute of Electrical and Electronics Engineers (IEEE) Transactions on Medical Imaging 29, 196–205.

Klikauer, T., 2016. Scikit-learn: Machine Learning in Python. TripleC 14, 260–264. URL: http://dl.acm.org/citation.cfm?id=2078195{%}5Cnhttp://arxiv.org/abs/1201.0490, doi:10.1007/s13398-014-0173-7.2, arXiv:arXiv:1011.1669v3.

Levi, C.R., Parsons, M.W., Christensen, S., Davis, S.M., Donnan, G.A., Desmond, P.M., Campbell, B.C., 2011. Cerebral Blood Flow Is the Optimal CT Perfusion Parameter for Assessing Infarct Core. Stroke 42, 3435–3440.

Liu, C., 2018. IVDM3Seg Challenge, in: International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2018 Challenge on Automatic IVD Localization and Segmentation from 3D Multi-modality MR (M3) Images.

Luna, M., Park, S.H., 2018. 3D Patchwise U-Net with Transition Layersfor MR Brain Segmentation, in: Grand Challenge on MR Brain Segmentation at International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) 2018.

Mahmood, Q., Basit, A., 2016. Segmentation of Ischemic Stroke Lesion using Random Forests in Multi-modal MRI Images, in: Ischemic Stroke Lesion Segmentation (ISLES) 2016. URL: http://www.isles-challenge.org/ISLES2016/pdf/20160927_ISLES2016_Proceedings.pdf.

Maier, O., Handels, H., 2016. Predicting Stroke Lesion and Clinical Outcome with Random Forests, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Second International Workshop BrainLes 2016, Springer International Publishing. pp. 218–230.

McKinney, W., 2011. pandas: a Foundational Python Library for Data Analysis and Statistics, in: Python for High Performance and Scientific Computing (PyHPC) 2011, pp. 1–9.

Meschia, J.F., Brott, T., 2018. Ischaemic stroke. European Journal of Neurology 25, 35–40.

Muir, K.W., Buchan, A., von Kummer, R., Rother, J., Baron, J.C., 2006. Imaging of acute stroke. The Lancet Neurology 5, 755–768.

Patel, S.C., Levine, S.R., Tilley, B.C., Grotta, J.C., Lu, M., Frankel, M., E. Clarke Haley, J., Brott, T.G., Broderick, J.P., Horowitz, S., Lyden, P.D., Lewandowski, C.A., Marler, J.R., Welch, K.M.A., rt PA Group, f.t.N.I.o.N.D., Stroke, Study, S., 2001. Lack of Clinical Significance of Early Ischemic Changes on Computed Tomography in Acute Stroke. JAMA 286, 2830.

Pfeiffer, F.W., 2007. Automatic differentiation in PyTorch, in: 31st Conference on Neural Information Processing Systems (NIPS 2017), pp. 2–8.

Powers, W.J., Rabinstein, A.A., Ackerson, T., Adeoye, O.M., Bambakidis, N.C., Becker, K., Biller, J., Brown, M., Demaerschalk, B.M., Hoh, B., Jauch, E.C., Kidwell, C.S., Leslie-Mazwi, T.M., Ovbiagele, B., Scott, P.A., Sheth, K.N., Souther-

land, A.M., Summers, D.V., Tirschwell, D.L., 2018. 2018 Guidelines for the Early Management of Patients With Acute Ischemic Stroke: A Guideline for Healthcare Professionals From the American Heart Association/American Stroke Association. volume 49. arXiv:1608.04207.

Rankin, J., 1957. Cerebral Vascular Accidents in Patients over the Age of 60: II. Prognosis. Scottish Medical Journal 2, 200–215.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. Lecture Notes in Computer Science 9351, 234–241. arXiv:1505.04597.

Rubinsteyn, A., Feldman, S., . fancyimpute 0.4.3. URL: https://pypi.org/project/fancyimpute/.

van Swieten, J.C., Koudstaal, P.J., Visser, M.C., Schouten, H.J., van Gijn, J., 1988. Interobserver agreement for the assessment of handicap in stroke patients. Stroke 19, 604–607.

Uchida, K., Yoshimura, S., Sakai, N., Yamagami, H., Morimoto, T., 2019. Sex Differences in Management and Outcomes of Acute Ischemic Stroke With Large Vessel Occlusion. Stroke , STROKEAHA119025344.

Vilela, P., Rowley, H.A., 2017. Brain ischemia: CT and MRI techniques in acute ischemic stroke. European Journal of Radiology 96, 162–172.

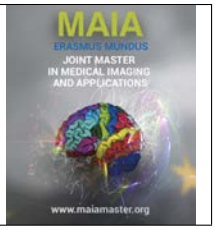
Vrtkova, A., 2017. Predicting clinical status of patients after an acute ischemic stroke using random forests. Proceedings of the International Conference on Information and Digital Technologies, IDT 2017 , 417–422.

Van der Walt, S., Schönberger, J.L., Nunez-Iglesias, J., Boulogne, F., Warner, J.D., Yager, N., Gouillart, E., Yu, T., 2014. scikit-image: image processing in python. PeerJ 2, e453.

Yushkevich, P.A., Piven, J., Hazlett, H.C., Smith, R.G., Ho, S., Gee, J.C., Gerig, G., 2006. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. NeuroImage 31, 1116–1128.

Medical Imaging and Applications

Master Thesis, June 2019



Soft tissue lesion detection in digital breast tomosynthesis using domain adaptation from mammograms

Mahlet Alie Birhanu[a], Mehmet Ufuk Dalmis[b], Michiel Kallenberg[b], Jaap Kroes[b]

[a] *Université de Bourgogne (France), UNICLAM (Italy) and Universitat de Girona (Spain)*
[b] *ScreenPoint Medical, Nijmegen, The Netherlands*

Abstract

Digital mammography (DM) has been a standard two-dimensional (2D) imaging modality for breast cancer screening for many decades. Although large clinical trials have shown that screening mammography improves early detection of cancer, mammographic sensitivity is shown to be lower due to the summation of overlapping breast tissue during image acquisition. For such reasons, digital breast tomosynthesis (DBT) is becoming a preferred technology used in diagnostic breast imaging as it guarantees improved visualization of breast detail. Deep learning based Computer-aided detection (CAD) has been used in screening digital mammography and is currently in use for DBT as it improves screening sensitivity and specificity. But given the fact that DBT is a relatively new technology and is not as widely used as mammography, it is difficult to collect a sufficient amount of malignant abnormalities to train a deep learning based CAD system. Therefore, in this work, we propose to use the recently developed domain adaptation methods to convert DBT images to synthetic DMs so that available CAD systems can be used to perform detection of soft tissue lesions (STL) on the synthesized DM images. A modified version of a framework called CycleGAN was used for unpaired-image-to-image translation. The dataset used for training, validation, and testing includes over 30,000 extracted patches from exams of 2 different vendors. A CNN model previously trained on real DM images was used to evaluate how well the synthetic images can be classified. Classification results of synthetic DM images on validation set showed a 8% improvement in AUC (AUC = 0.90) compared to a baseline performance of real DBT classification with an AUC = 0.82. An AUC of 0.88 was achieved on a separately held testing set with a baseline DBT classification of AUC = 0.80. The results indicate the ability of the modified framework on generating realistic synthetic images.

Keywords: Digital breast tomosynthesis (DBT), CycleGAN, Soft tissue lesion (STL), GAN, Digital mamography(DM)

1. Introduction

Breast cancer is the second most common cause of cancer death in women (Siegel et al., 2018). Mortality can be reduced by 30% using digital mammography screening (Bazzocchi et al., 2007).

Digital mammography (DM) has been a standard two-dimensional (2D) imaging modality for breast cancer screening for many decades. Mammograms are relatively low-dose soft tissue X-ray images of the breast. Conventionally, both left and right breasts are imaged using two standard views, the cranial-caudal (CC) and the mediolateral-oblique (MLO). The resultant 2D image maps the integral x-ray attenuation of tissues in a single plane. Large clinical trials have shown that screening mammography improves early detection of cancer and increases survival (Broeders et al., 2012), (Lauby-Secretan et al., 2015). Screening mammography has a sensitivity of ~ 70% in women 49-69 years old, which means that 30% of tumours are missed on mammography (Warren Burhenne et al., 2000a). Mammogram sensitivity is even lower in young women and women with dense breasts. Therefore, an improvement in sensitivity can help to significantly reduce mortality.

Digital breast tomosynthesis (DBT) is an emerging technology used in diagnostic breast imaging to evaluate potential abnormalities and is being more preferable than digital mammography as the standard x-ray technique for breast cancer screening. In DBT,

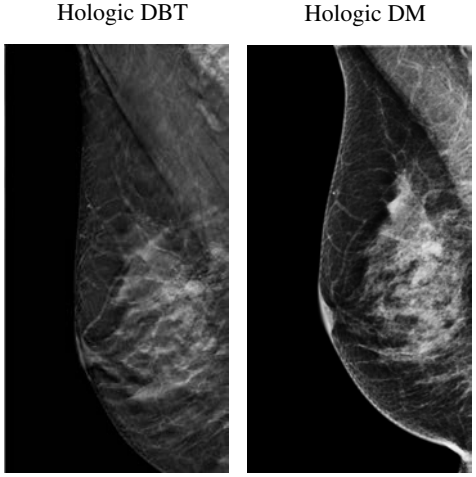


Figure 1: MLO view of a breast slice in tomosynthesis(a) and mammogram(b)

the compressed breast tissue is imaged in a quasi-three-dimensional manner by performing a series of low-dose radiographic exposures and using the resultant projection image dataset to reconstruct cross-sectional in-plane images in standard mammography views (M Hakim et al., 2010). Generally, a series of projection view images is acquired as the x-ray source is rotated about the fulcrum over a limited range of angles. Because of the wide dynamic range and high detection ability of digital detectors, each of the projection images can be acquired with a fraction of the x-ray exposure used for a regular mammogram. The total dose required for DBT may therefore be kept at nearly the same or slightly higher than that of a regular mammogram. Tomographic slices focused at any depths of the imaged volume can be generated with reconstruction techniques from the series of projection images (Chan et al., 2008).

Improved visualization of breast detail in DBT allows improved characterization of findings, including normal structures and breast cancer. This technology reduces the summation of overlapping breast tissue, which can cover up/hide breast cancer, and provides improved detail of non-calcified findings seen in breast cancer. It also assists in lesion localization and determining the extent of lesions in women with known or suspected breast cancer (M Hakim et al., 2010). Much of the improvement in screening outcomes achieved with DBT is the result of better differentiation between overlapping glandular tissue and subtle asymmetries, distortions, and margin characterization of masses (P. Zuckerman et al., 2017). By scrolling through the DBT slices, readers are able to more confidently differentiate between a discrete lesion and overlapping normal glandular tissues. For such reasons, studies of DBT in the screening setting have reported decreased recall rates (improved specificity) or increased cancer detection rates (improved sensitivity) or both compared with standard 2D full-field digital mammography (FFDM) (Giess et al., 2017).

To further improve sensitivity and specificity of breast cancer detection, Computer-aided detection (CAD) (Warren Burhenne et al., 2000b) has been used in screening digital mammography for many years and is currently being developed to be utilized for DBT. Current CAD systems heavily rely on deep learning algorithms that aim to generalize a solution for unseen data based on the given training data. For the development of such systems, a large database containing training samples is of paramount importance. Yet, given the fact that DBT is a relatively new technology and is not as widely used as mammography, it is difficult to collect a sufficient amount of malignant abnormalities.

Since there is a significantly large database of digital mammography images, it could be of great value to use the available database to train DBT CAD systems. Using information from one domain (e.g. 2D DM) that is closely related to another the domain of interest (e.g. DBT) is generally known as transfer learning. A common approach in transfer learning is to utilize samples from the related domain for pretraining of a classifier. The work of Samala et al. (2016), for instance, demonstrates this approach by pretraining the first layers of a convolutional neural network (CNN) with DM images, and then using DBT data to finetune the last layers. A different approach would be to transfer the samples from a source domain to the target domain. Recently, in deep learning, generative models have been proposed to go from one domain to the other. Generative adversarial networks (GAN) have been used in many different settings for a wide range of applications, including collection style and domain transfer, object transfiguration, season transfer and photo enhancement (Goodfellow et al., 2014).

Due to the impressive results GANs produced, they are now also being used in paired image-to-image translation for synthesizing medical images. Although the goal of image-to-image translation is to learn the mapping between a source image and a target image using a training set of aligned image pairs, paired training data will not be available for many tasks. For this reason, Zhu et al. (2017) presented an approach for learning to translate an image from a source domain to a target domain in the absence of paired examples. This approach, referred to as CycleGAN is more practical in medical images, as a big dataset of paired images from different modalities is usually not available.

The purpose of this study is to use the same concept of domain adaptation from DBT images to DM so that existing CAD systems can be used to perform detection of soft tissue lesions (STL) from synthetic mammograms. The DBT images look visually different from FFDM images as can be seen in Figure 1. By the use of GANs, DBT image slices can be used to synthesize "mammography looking" images that can serve as a direct input to the available CAD systems trained on real

DM. In principle, the resulting synthetic images should have the same appearance and discriminating features as real DMs when seen by a classifier to detect soft tissue lesions from normal ones. Feeding the synthetic images to a STL classifier trained on real DM images can serve as a good way of evaluating how similar the generated images are to the real ones.

2. State of the art

2.1. GAN

The framework for training generative models in an adversarial manner for synthetic image generation was first introduced in the seminal work of Goodfellow et al. (2014). This framework is based on a simple but powerful idea: images are sampled from a simple distribution (e.g. a normalized Gaussian) in a low-dimensional space known as latent space. Each latent vector in this latent space is transformed into an image using a generator neural network. This generator neural network aims to produce realistic examples able to deceive the discriminator which aims to discern between original and generated ones. The two networks form an adversarial relationship and gradually improve one another through competition, much like two opponents in a zero-sum game.

Various flavors of GANs have been recently proposed, both purely unsupervised as shown in the works of Arjovsky et al. (2017) and Berthelot et al. (2017) as well as conditional. Many different flavors of GAN optimization problems differ by the constraint on the discriminators output and corresponding loss, and the presence and application of gradient norm penalty. While these models achieve compelling results in specific domains and were aimed at improving the overall performance of GANs, there is still no clear consensus on which GAN algorithm(s) perform objectively better than others (Lucic et al., 2018). This is partially due to the lack of robust and consistent metrics. According to the research conducted by Lucic et al. (2018), there was no clear evidence that any of the tested modified GAN algorithms consistently outperforms the original GAN introduced in the work of Goodfellow et al. (2014).

The work of research by Lucic et al. (2018) provided a comprehensive comparison of the state-of-the-art GANs used for image generation, and empirically demonstrate that nearly all of them can reach similar values of Fréchet Inception Distance (FID) (Heusel et al., 2017). But all these evaluations were based on GANs performing image generation tasks rather than domain transfers.

2.2. CycleGAN

Since paired images in different domains are not usually available for training in many problems, Zhu et al. (2017) presented their work on unpaired image-to-image translation based on the working principles of GANs for image generation. According to their work, in the presence of unpaired images from different domains, the same concept of "back translation and reconciliation" used in human translators can be applied to images as well. To this end, they used two GANs in a cycle, by objectively trying to minimize cycle consistency loss. The GANs are used to generate target image from source (or vice versa) and cycle consistency makes sure that each generated image can be reconstructed back to the original image with a minimal loss of information. This technique reduces the probability of generating an image that can not be distinguished from a real image in the target domain but entirely deviating from the structure of the source image. CycleGAN has been used in a Cityscapes dataset to convert semantic label to a photo (and vice verca) as well as map to a real photo conversion.

Although cycle consistency loss is the main reason why CycleGAN works in translating unpaired imaging data, it is also found to give the model an intriguing property of "hiding" information about a source image into the images it generates in a nearly imperceptible, high frequency signal as shown by the work of Chu et al. (2017). This unpredicted property ensures that the generator can recover the original sample and thus satisfy the cyclic consistency requirement, while the generated image remains realistic.

2.2.1. CycleGAN with Better Cycles

In order to alleviate the previously mentioned problematic nature of cycle consistency, Wang and Lin (2017) propose simple modifications to CycleGAN, and showed that such an approach achieves better results. Some of the proposed changes are, decaying the weight of cycle consistency loss as training progresses and additionally weighting cycle consistency loss by the quality of generated images. The modification was justified by only visual inspection and comparison with the original CycleGAN implementation.

2.2.2. Conditional CycleGAN

In the work of Mirza and Osindero (2014) and Odena et al. (2017), GANs were extended to a conditional model by conditioning both the generator and the discriminator on some extra information such as class labels or data from other modalities. This method provided an alternative to have more control of the generation of images according to a flexible external information. Promising results were obtained in face generation and swapping. However, this method is not applicable for unlabeled data (for example in a classification problem in deployment).

Table 1: Summary of published applications of GANs in medical imaging used for unpaired image to image translation tasks

Reference	Medical image type	Resolution	# of samples	Algorithm	Assessment	Quantitive metric
Wolterink et al. (2017)	MRI/CT	183× 288× 288	24	CycleGAN	Visual	MAE
Becker et al. (2018)	DM	256×256 512×408	680	CycleGAN	Radiologist readout	AUC
Seeböck et al. (2019)	OCT	496×512×49 1024×512×128 1024×200×200	1407	CycleGAN	Segmentation	F1-score
Armanious et al. (2019)	PET/CT	256×256	2355	Cycle-MedGAN	Visual	PSNR SSIM LPIPS VIF
Korkinof et al. (2018)	DM	1280×1024	1000K	GAN	Visual	-

Note: Mean Absolute Error (MAE), Area Under Curve (AUC), Peak Signal to Noise Ratio (PSNR), Structural Similarity Index (SSIM), Learned Perceptual Image Patch Similarity (LPIPS), Visual Information Fidelity (VIF)

Table 2: Dataset A: Distribution of dataset used for training and validation

Patch type	Abnormal		Normal	
	Training	Val	Training	Val
Hologic DM	6446	350	9000	1000
Siemens DBT	500	79	9000	1000

Table 3: Dataset A sub-sampled: Distribution of balanced dataset used for training and validation

Patch type	Abnormal		Normal	
	Training	Val	Training	Val
Hologic DM	6446	350	6448	1000
Siemens DBT	500	79	500	1000

Table 4: Clinical dataset distribution: independent dataset used for testing

Patch type	Abnormal	Normal
Hologic DM	226	1253
Siemens DBT	228	1262

2.3. UNIT

UNsupervised Image-to-image Translation method (UNIT) aims at learning a joint distribution of images in different domains by using images from the marginal distributions in individual domains. As shown by research conducted by (Liu et al., 2017), this method is based on GANs and takes a shared-latent space assumption to alleviate the existence of an infinite set of joint distributions arriving at the given marginal distributions.

2.4. GANs in medical images

Although we have not come across a research paper directly dealing with domain adaptation from DBT to 2D DM data, there are plenty of reported results of GANs used in medical images. Table 1 summarizes the use of GANs in multiple imaging modalities.

3. Material and methods

3.1. Dataset

Exams in the training and testing sets were collected at multiple clinical centers across Europe, including sites in the Netherlands, Germany, and the UK. Data collection sites are representative for regular breast cancer screening and asymptomatic patients in hospitals who have mammograms for a variety of reasons, such as increased risk for breast cancers or not being invited for population-based screening program, e.g. because of age under 50. For the inclusion of the normal exams in the test set, a follow-up of at least one year was required. Most of the exams in the test set have MLO and CC views of both the left and right breast. Patch extraction from a normal exam results in approximately 60 candidate patches (15 per image), while an abnormal exam containing a soft tissue lesion typically gives about 2 abnormal patches. For DBT exams, patches were extracted from a volume. In total, more than 180,000 exams were used for patch extraction. Once all patches were extracted, randomly chosen patches were used for training and validation whose distribution is shown in

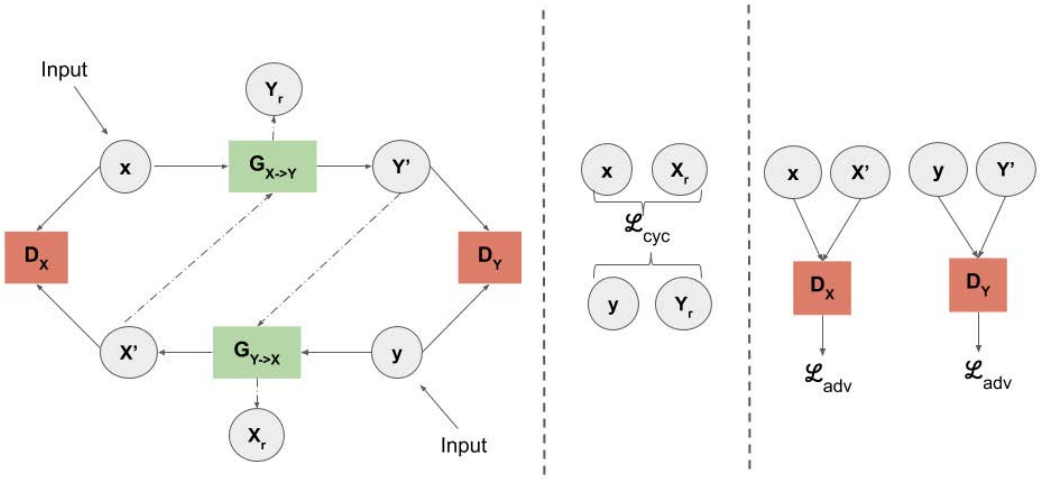


Figure 2: An overview of the CycleGAN framework for unpaired image translation. x and y are unpaired images randomly sampled from their respective domains. X' and Y' are synthetic images created by generators and X_r and Y_r are reconstructed images.

Table 2 and 3. Truthing was based on the clinical reports from both radiology and pathology. A region was considered a true positive if the center of the patch falls within the annotated contour of the lesion. Training and validation split was done on patch level. All DBT patches are extracted from Siemens DBT slices and DM patches from Hologic images as shown in Table 2, 3 and 4.

A subset of Dataset A described in Table 2 was taken with balanced number of normals and abnormals to train modified CycleGAN with a weight of 1. Details are shown in Table 3. Additionally, a separately held testing set was used, details of which are shown in Table 4.

3.2. Preprossessing

All images were preprocessed by applying window-level followed by energy band normalization and center cropping prior to patch extraction in the same manner as in the work of Kooi et al. (2017) . Patches were given in a 16-bit format. Additionally, CycleGAN output patches needed to be center cropped to a size of 224×224 pixels to be fed to the CNN classifier.

3.3. The CycleGAN framework

A Generative Adversarial Network (Goodfellow et al., 2014) consists of two neural networks, a generator $G_{X\to Y}$ and a discriminator D_Y, which are iteratively trained in a two-player minimax game manner. The adversarial loss $\mathcal{L}_{adv}$ ($G_{X\to Y}$, D_Y) is defined as,

$$\mathcal{L}_{adv}(G_{X\to Y}, D_Y) = \min_{\theta_g}\max_{\theta_d}\{\mathcal{E}[D_Y(y)] + \mathcal{E}[(1 - D_Y(G_{X\to Y}(x))]\} \quad (1)$$

where θ_g and θ_d are respectively the parameters of the generator $G_{X\to Y}$ and discriminator D_Y , and $x \in X$ and $y \in Y$ denotes the unpaired training data in source and target domain respectively. $\mathcal{E}$ represents the loss function used to evaluate generator and discriminator outputs. $\mathcal{L}(G_{Y\to X}$, D_X) is analogously defined. In CycleGAN, X and Y are two different image representations, DBT and DM this case and the CycleGAN learns the translation $X \to Y$ and $Y \to X$ simultaneously. Training data in CycleGAN is unpaired. Training the framework merely with the adversarial losses is not sufficient since it may lead to mode collapse, where a set of different input images are mapped into a single image in the target domain (Zhu et al., 2017). Therefore, an additional constraint regularizing the mapping functions is essential. This was achieved by introducing Cycle Consistency to enforce forward-backward consistency which can be considered as pseudo pairs of training data. Cycle consistency loss used in CycleGAN is defined as,

$$\mathcal{L}_{cyc}(G_{X\to Y}, G_{Y\to X}) = \|G_{Y\to X}(G_{X\to Y}(x)) - x\| + \|G_{X\to Y}(G_{Y\to X}(y)) - y\| \quad (2)$$

The total loss function of CycleGAN will consequently be:

$$\mathcal{L}(G_{X\to Y}, G_{Y\to X}, D_X, D_Y) = \mathcal{L}_{adv}(G_{X\to Y}, D_Y) + \mathcal{L}_{adv}(G_{Y\to X}, D_X) + \lambda\mathcal{L}_{cyc}(G_{X\to Y}, G_{Y\to X}) \quad (3)$$

The cycle consistency loss is weighted by some λ to have more control of its effect on the reconstructed images. A flowchart that depicts an overview of the CycleGAN framework is shown in Figure 2. The generator network used contains two stride-2 convolutions, several residual blocks, and two fractionally strided convolutions with stride $1/2$. 9 blocks were used input image

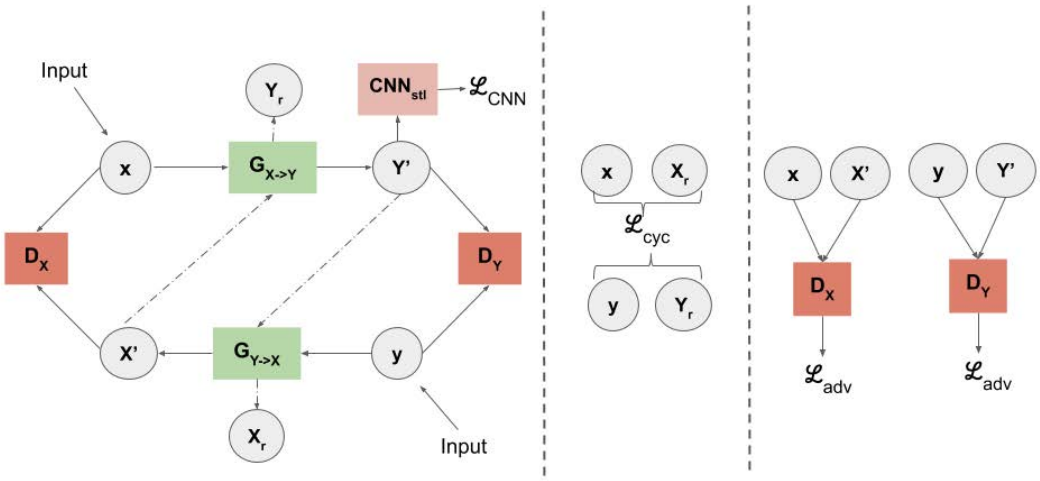


Figure 3: An overview of the proposed CycleGAN framework: CNN_{stl} is introduced to guide generator to output correctly classifiable results.

resolutions higher than 256 256 pixels as shown by Zhu et al. (2017). For the discriminators, a PatchGAN based network ((Isola et al., 2016)) was used as in the original implementation. Such a patch-level discriminator architecture has fewer parameters than a full-image discriminator and can work on arbitrarily-sized images in a fully convolutional fashion aim to classify whether overlapping image patches are real or fake as implemented by Zhu et al. (2017) and Isola et al. (2016).

3.4. Proposed framework

Although CycleGAN relies on the cycle-consistency loss to avoid mismatches which could occur due to unsupervised training using unpaired images, it has been discussed in the work of Chu et al. (2017) that it can cause the model to hide information about a source image into the images it generates. In this way, although the translated DBT images resemble real DM images visually, they could still maintain information from the source images. Additionally, since our data comprise of patches from two classes, the same architecture can not benefit from such information. To circumvent this, we propose to introduce an additional loss function, $\mathcal{L}_{CNN}$, that can condition CycleGAN on how well the generated DM images can be correctly classified with a pretrained classification network (CNN_{stl}). Much like the implementation of conditional CycleGANs (Mirza and Osindero, 2014), the loss is conditioned on an external information, which is a class label in this case. This classifier network is pretrained on real DM images to classify between images containing soft tissue lesions and normal ones. As the training set consists of lesions patches and normal ones, this loss function can guarantee that a lesion containing DBT image patch gets transformed into a lesion containing DM and the same holds true for normal candidates. $\mathcal{L}_{CNN}$ is defined as,

$$\mathcal{L}_{CNN}(G_{X\to Y,CNN_{stl}}) = \lambda_c \min_{\theta_g}\{\mathcal{E}(CNN_{stl}(G_{X\to Y}(x)))\} \tag{4}$$

In this new implementation of CycleGAN, the same concept is used for the other losses, i.e. $\mathcal{L}_{adv}$ and $\mathcal{L}_{cyc}$. In principle, $\mathcal{L}_{CNN}$ is expected to guide the generator to output correctly classifiable synthetic DM images. Hence, $\mathcal{L}_{CNN}$ is only used to evaluate the generated synthetic images and the CNN model does not get altered in the process. The overall CycleGAN loss becomes,

$$\begin{aligned}\mathcal{L}(G_{X\to Y}, G_{Y\to X}, D_X, D_Y) = \mathcal{L}_{adv}(G_{X\to Y}, D_Y) \\ +\mathcal{L}_{adv}(G_{Y\to X}, D_X) + \lambda\mathcal{L}_{cyc}(G_{X\to Y}, G_{Y\to X})+ \\ \lambda_c\mathcal{L}_{CNN}(G_{X\to Y})\end{aligned} \tag{5}$$

A summary of the proposed framework is shown in Figure 3.

3.4.1. CNN architecture

The CNN_{stl} network has a VGG-11 architecture as shown in Figure 4. It consists of 11 convolutional and two fully connected layers with ReLu activation. Batch normalization was also used. The network was trained with real DM images. Input patches were downsampled from the original resolution to a pixel-spacing of 200 microns cropped to 224×224 pixels. The final output was determined with a softmax layer.

3.5. UNIT

Since the working principle of UNIT is entirely different from CycleGAN, our dataset was trained using the original open source implementation of UNIT and results were used for comparison purposes.

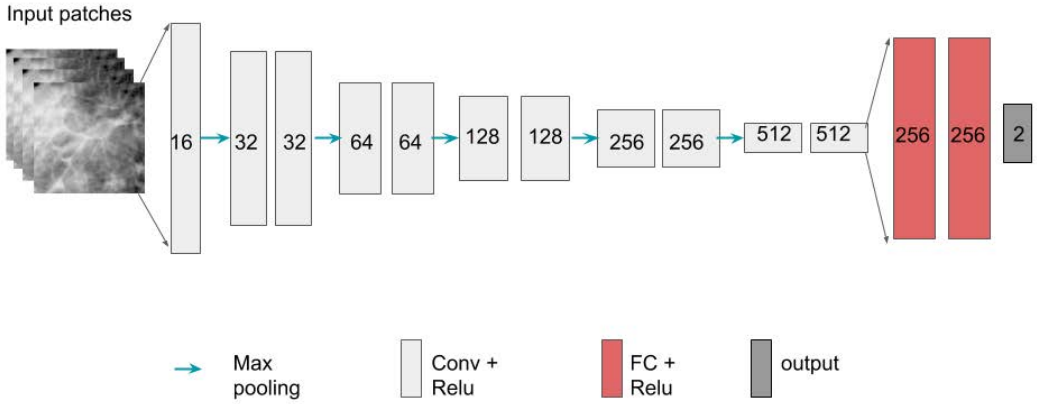


Figure 4: Illustration of CNN_{stl} model architecture: Input patches of 224×224 are given to a VGG-11 architecture model with 2 fully connected (FC) layers

3.6. Evaluation metric

Besides visual inspection, ROC analysis of the CNN_{stl} classifier was used to quantitatively evaluate how well the generated synthetic DM images resemble real DM images. Classifier performance on a testing set of real DM images was used as an upper bound and AUC from classification of real DBT images with the same classifier was used as a lower bound. The DBT and DM images in the testing set are unpaired. In principle, the generated images should look like a real DM and have classification result that is higher than the lower bound and somewhat approaching the upper bound limit.

4. Results

4.1. Training details

Several experiments were run to evaluate the plain and modified CycleGAN in the task of unpaired DBT-to-DM translation. For all experiments, the Adam optimizer was used with the parameters $\beta_1 = 0.5$ and $\beta_2 = 0.9$ as recommended by the original CycleGAN implementation of Zhu et al. (2017). The learning rate was kept constant for the first 100 epochs, and for the rest of the epochs, it is linearly decayed to 0. The default value was used for the weight of the cycle consistency loss ($\lambda = 10$). When training the discriminator, the loss was divided by 2 before back-propagating. The weights were initialized with a Gaussian distribution with a mean 0 and a standard deviation of 0.02. The learning rate (η), batch_size and the weight of CNN loss (λ_c) were tuned differently for different experiments as shown in Table 5. Training progress of both generators and discriminators was tracked in all experiments to ensure that the model is actually learning. Experiments were terminated when the generator and discriminator loss seem to show no further improvement.

4.2. Baseline and upper bound

The baseline of all experiments was taken to be the performance of CNN_{stl} classifier on real DBT patches both in the validation and testing set. The baseline AUC was evaluated to be 0.82 on the validation set and 0.80 on the testing set as shown in Table 5. The same patches used for the baseline were transformed by models from the experiments and their classification result is compared to the baseline. As an additional measure of comparison, the performance of the same classifier was evaluated on real DM patches in the testing set. AUC on real DM patches is taken to be an upper-bound estimate. Although these patches are a different set of patches, it helps to give an idea of how real DM images can be classified. Upper-bound AUC is calculated to be 0.90 in the testing set.

4.3. Plain CycleGAN

To have a preliminary understanding of how well the original CycleGAN implementation performs on the task of unpaired DBT-to-DM image translation, multiple experiments were carried out. For both domains, the number of normal samples was a lot higher than the abnormals. In order to avoid random pairing of an abnormal sample with a normal one during training, samples from only one class (normal or abnormal) were exclusively used for training in the first phase of experiments. The resulting model was then used for translation on the

No.	Algorithm	Patches	η	λ_c	BS	Norm.	Best epoch	Total epoch	AUC	
									Val.	Test
Baseline	-	-	-	-	-	-	-	-	0.82	0.80
1	CycleGAN	normals	0.0002	-	1	instance	30	50	0.78	0.76
2	CycleGAN	normals	0.0001	-	4	batch	50	60	0.80	0.79
3	CycleGAN	abnormals	0.0001	-	4	instance	50	200	0.78	0.76
4	CycleGAN	abnormals	0.0002	-	1	instance	80	100	0.84	0.81
5	CycleGAN	abnormals + normals	0.0002	-	1	instance	90	95	0.81	0.78
6	CycleGAN	abnormals + normals (balanced)	0.0002	-	1	instance	90	100	0.81	0.78
7	CycleGAN + CNN	abnormals + normals	0.0002	18	1	instance	110	130	**0.90**	**0.88**
8	CycleGAN + CNN	abnormals + normals (balanced)	0.0002	1	1	instance	70	150	**0.90**	0.85

Table 5: Summary of results on selected experiments; η refers to the learning rate of CycleGAN, λ_c refers to the weight of CNN loss for abnormal samples. BS refers to the batch size used.

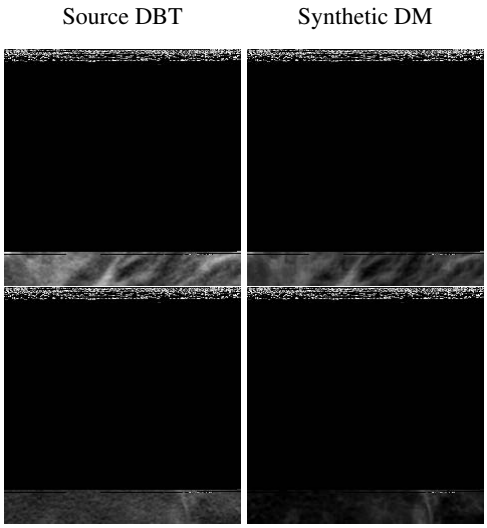


Figure 5: An illustrative example of CycleGAN trained with only normal patches: translation result of lesion patches

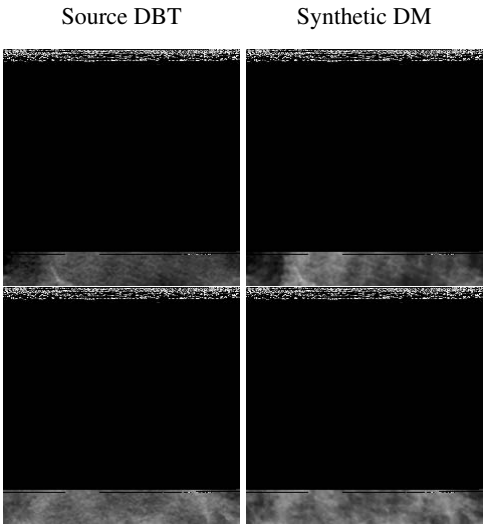


Figure 6: An illustrative example of CycleGAN trained with only abnormal patches: translation result of normal patches

validation and testing sets containing patches from both classes. Experiments were repeated by adjusting hyperparameters (batch size and learning rate) to improve results. Additionally, another phase of experiments was carried out by using samples from both classes in the training set to see the effect of random pairing.

4.3.1. CycleGAN on normal patches

The original implementation of CycleGAN was trained with only normal patches in the training set. The resulting synthetic DM patches were evaluated with CNN_{stl} giving an AUC score of 0.80 in the best performing epoch model on the validation set and 0.79

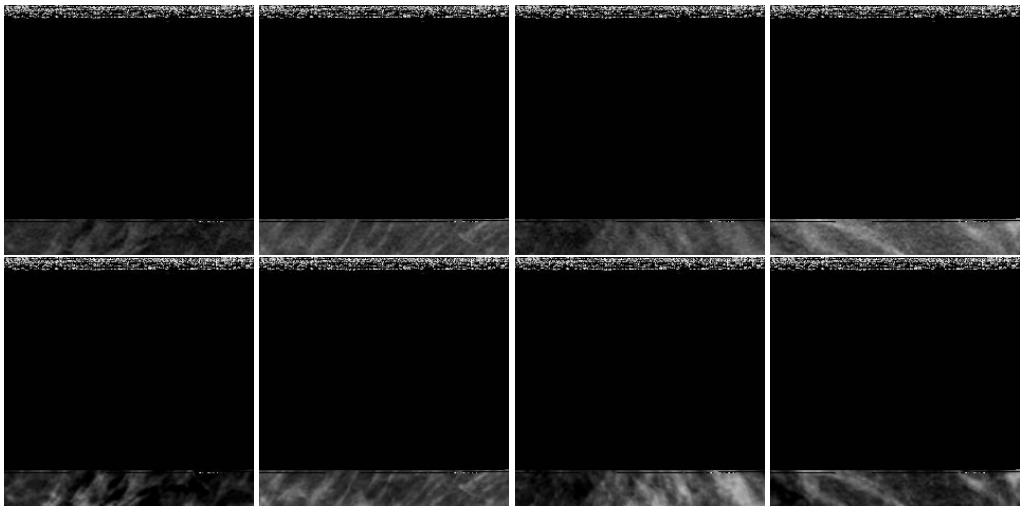


Figure 7: Illustrated example translation results of modified CycleGAN on abnormal samples: Patches on the top row are source DBT patches and on the bottom row are the translated DM results.

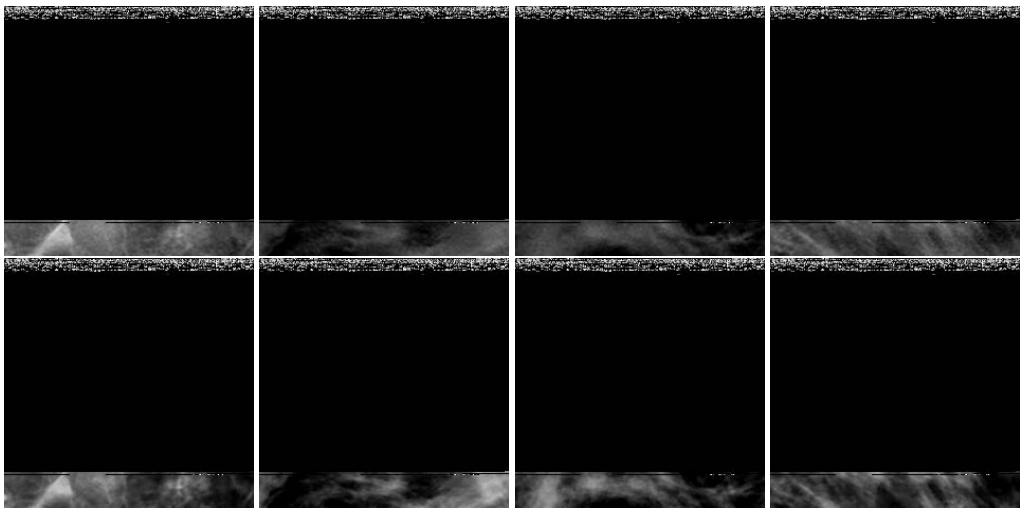


Figure 8: Illustrative example of translation results of modified CycleGAN on normal samples: patches on the top row are source DBT patches and on the bottom row are the translated DM results.

on the testing set. As compared to the baseline AUC, training CycleGAN with normal samples did not seem to improve the results. Some of the top misclassified abnormal images at a false positive rate (FPR) of 0.1 are shown in Figure 5. The qualitative results show that translated lesion patches seem to have a lower contrast although they pertain a visual characteristic different from a real DBT. The experiment was repeated with slightly adjusted hyperparameters (decreasing the learning rate and increasing the batch size). As shown in Table 5, none of the modifications seem to have helped boost the result.

4.3.2. CycleGAN on abnormal patches

Training CycleGAN with only abnormal patches was found to have an AUC score of 0.84 evaluated with the CNN_{stl} classifier. The best performing epoch model also showed a slight improvement ($\sim$ 1%) in the testing set giving an AUC of 0.81 as shown in Table 5. Although the result showd improvements, some of the top misclassified normal patches in this training at a FPR rate of 0.1 as shown in Figure 6 are found to have a very high contrast in most parts of the patch. Non-lesion areas look highlighted and the patches resemble abnormal patches in some parts.

4.3.3. CyceGAN on combined patches

The original implementation of CycleGAN was trained on patches from both classes. For this experiment, a CycleGAN model trained with only normal samples until the 50^{th} epoch was used as the starting pretrained model. The translated images in the validation set resulted in an AUC of 0.80 on the best epoch model showing lower performance compared to the baseline. As shown in Figure 10, all the training epoch models fall below the curve of the baseline in the

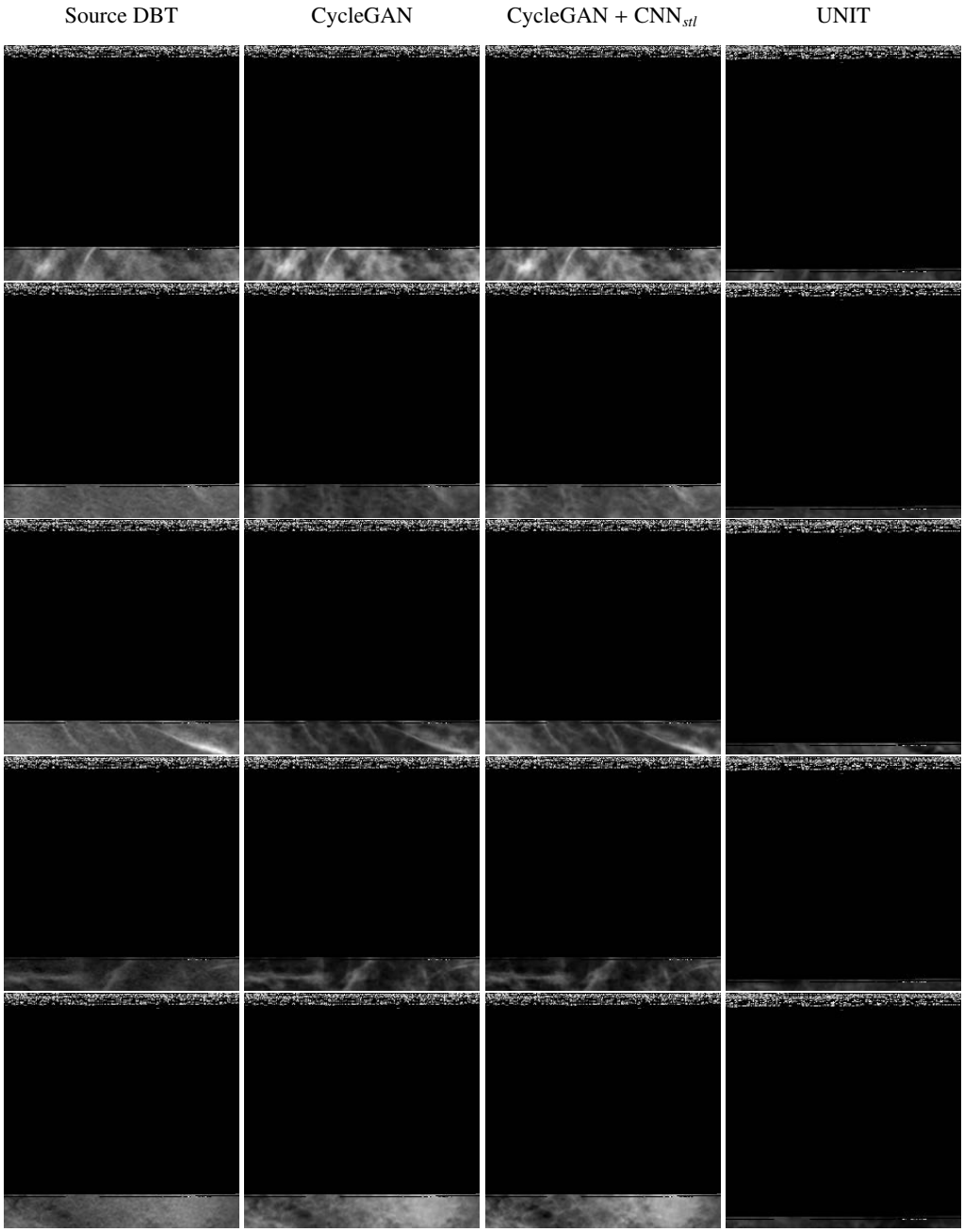


Figure 9: Comparison of unpaired DBT-to-DM translation results using different algorithms, i.e. CycleGAN, CycleGAN + CNN$_{stl}$ (modified CycleGAN) and UNIT. Left column shows images from the source DBT patches.

validation set.

4.4. Modified CycleGAN

Once the preliminary results were obtained from training CycleGAN with different subsets of the training set, the last set of experiments were staged to prove if the modified CycleGAN implementation can give a better result on the same task.

The proposed modified CycleGAN was also trained in the same setting on a dataset containing both normal and abnormal patches and starting from a pretrained model. During training, tracks of generator, discriminator and CNN losses were recorded. As shown in Figure 14, the CNN loss kept decreasing as the classification accuracy of generated samples decreased as the training progressed. Additionally, the differences in the losses between the plain and modified CycleGAN training were also recorded as shown in Figure 15. Experiments with modified CycleGAN resulted in the best improvement in AUC as shown in Figure 11. Visually, the translated images look very much like a real DM in both abnormal and normal lesions. For abnormal patches shown in Figure 7, lesion areas have a much

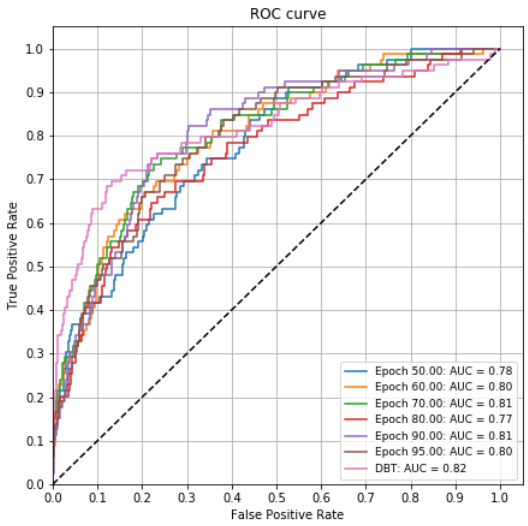


Figure 10: AUC results on validation set in different epochs of training CycleGAN with combined dataset

higher pixel intensity than surrounding areas much like a standard DM. The contrast of normal patches also resembles a real DM normal patch as seen in Figure 8. The modified CycleGAN weighted with $\lambda = 18$ to account for class imbalance on the training set, resulted in an AUC of 0.90 (8% improvement) on the validation set as shown in Figure 12 and 0.88 on the testing set showing a 8% improvement from the baseline and closely approaching the upper-bound performance as shown in Figure 13. The 95% confidence interval was calculated to be [0.850 - 0.937] on the validation set. Bootstrapped AUC result on the testing set along with the confidence interval area is shown in Figure 13. It can be seen that, the upper-bound AUC is within the confidence interval of the model and the baseline AUC falls below indicating an improved performance. Results of the proposed CycleGAN are presented in Figure 9 in comparison with UNIT and the original CycleGAN. In contrast, the resultant images produced by the original and modified CycleGAN framework have a global structure which closely resembles a real DM image. However, finer details are not highlighted when translated by CycleGAN such as speculations around a lesion. Moreover, the overall translated image in the modified CycleGAN seems to have better contrast between the background and tissues.

4.5. CycleGAN Vs UNIT

When training UNIT, default parameters were used as used in the work of Liu et al. (2017). The experiment was run for 700,000 iterations and was stopped after monitoring the results using the chosen metric. Qualitatively, the worse performance is exhibited by the UNIT framework. This is also reflected quantitatively, with UNIT resulting in the worst scores and the highest gap in confidence interval as shown in the bar graph 11 across the chosen metrics.

Overall, Figure 11 shows the performance of chosen experiments from each training scheme along with their confidence intervals evaluated in the validation set.

5. Discussion

In this work, we presented a modified unpaired image-to-image translation of DBT to DM images based on the concept of conditional CycleGANs as explained in Section 2 and 3. As a preliminary setup, the performance of original CycleGAN was evaluated on different subsets of the training set. As CycleGAN is a specific implementation of GANs to perform unpaired image-to-image translation, it learns how image from the source domain can be translated to an image in the target domain by randomly pairing images. Normally, this implementation is found to work quite well when the images in the target/source domains are essentially from one class. The intuition behind training CycleGAN with only normal/abnormal samples from the training set is to avoid the random pairing of an abnormal sample in the source domain to a normal one in the target domain (or vice versa.) However, translation results from CycleGAN trained with only normal samples showed to have a lower contrast and quality. Hence, abnormal samples in the validation set are translated to patches that are easily mistaken for a normal patch. Lesion areas and the surrounding speculations also look faded or non-existing. The underlying reason for such an effect might be because most of the normal patches in the training set comprise of samples that do not look like lesions and possibly because the average pixel intensity values of most areas in the patch are never as high as an area with a lesion. Most of the translated images look dark and many abnormal samples are missclassified.

Training CycleGAN with only abnormal patches had slightly better results. The generated synthetic patches look well contrasted. The reason for this could be because abnormal patches contain regions that can represent both normal and lesion containing samples. In such a way, the model learns to translate both abnormal and normal samples in a better way than the previous setup. Although the improvement is not much, it performs slightly better than the baseline as seen from Table 5 . From Figure 6, it can be seen that some normal patches are translated with more contrast in some regions than needed. This may have confused the classifier to recognize normal samples as abnormals, thereby not increasing its performance as needed. Training CycleGAN with both samples combined also did not seem to improve the results in any way. One good reason why this did not work as well could be because of the property of cycle-consistency loss. As mentioned in section

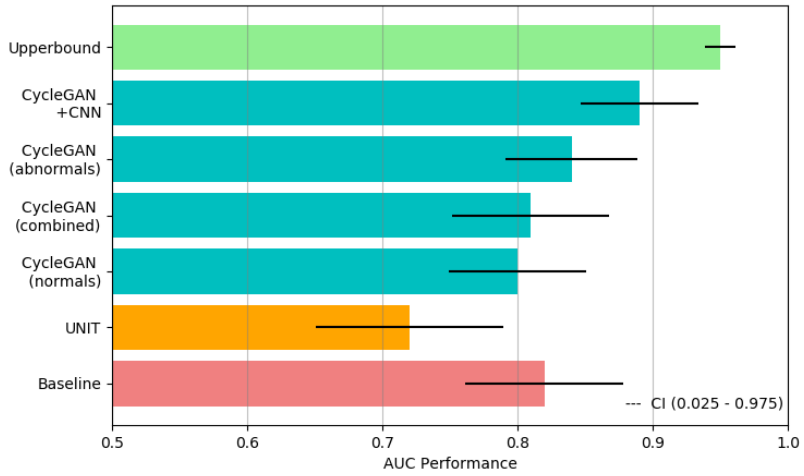


Figure 11: Comparison of AUC performance and 95% confidence interval of training models on validation set

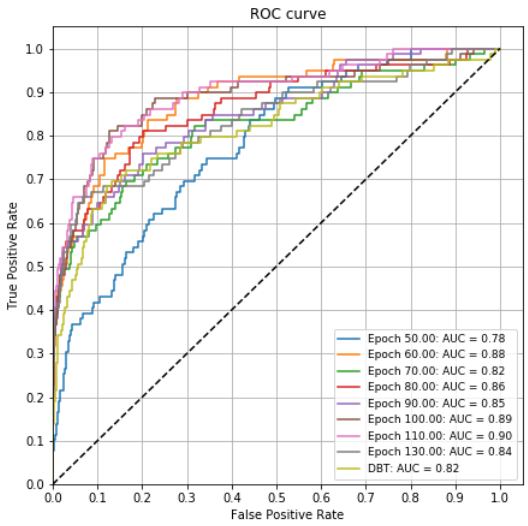


Figure 12: AUC results on validation set in different epochs of training CycleGAN + CNN_{stl}(modified CycleGAN) with combined dataset

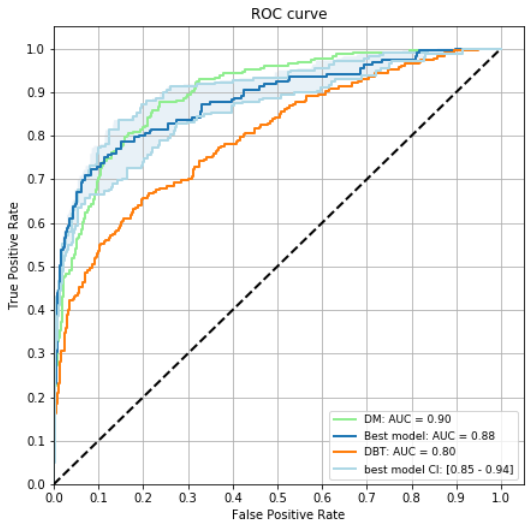
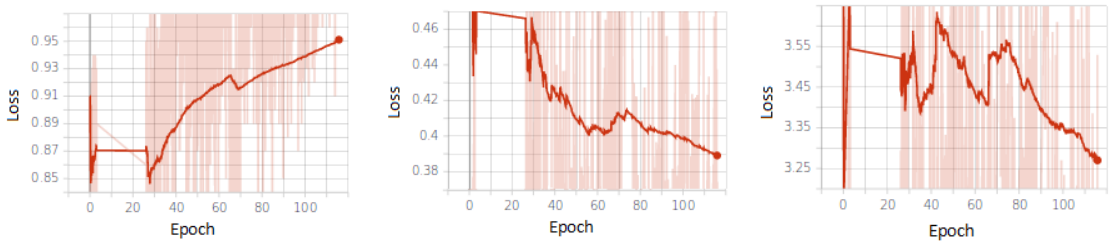


Figure 13: Best AUC results of CycleGAN + CNN_{stl} on testing set

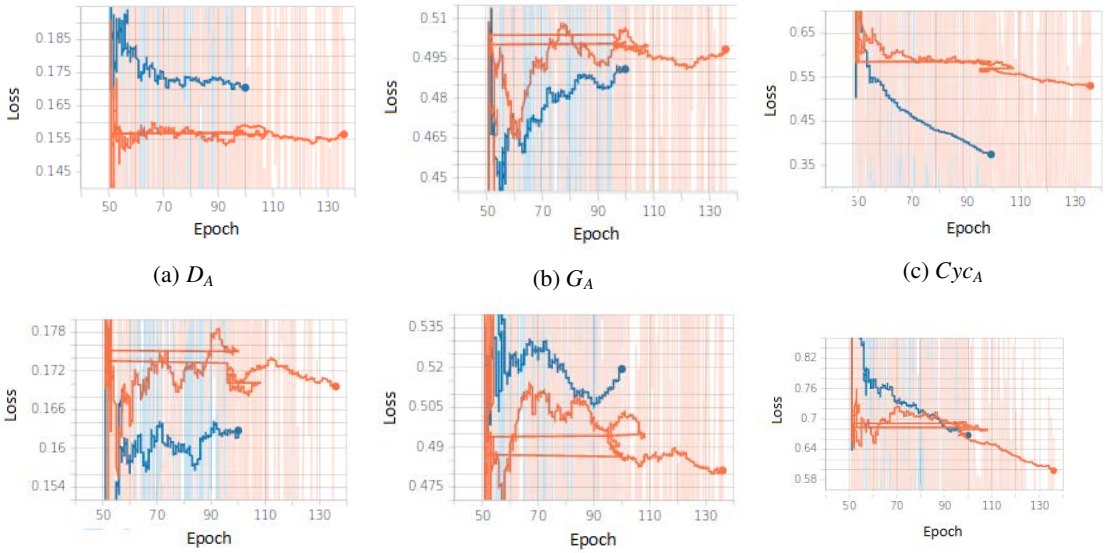
2, although cycle-consistency loss is the reason why CycleGAN works, it also has the disadvantage of hiding information from the source domain in order to do a good job during reconstructing back to the source domain. This is a logical thing to do for the network to keep decreasing the cycle-consistency and the overall loss. However, this could also hinder the CNN_{stl} classifier from detecting the important features to look for lesions. Hence, it is logical to explain that it performs just as good as the baseline (CNN_{stl} performance on real DBT) since hidden properties of the source images (DBT) could still be present in the translated results. But this hypothesis has not been tested and future work will focus on it.

The hypothesis behind using the modified CycleGAN implementation is to give the generators some conditional information about the class of the patches. In principle, the network should start producing patches that can be correctly classified by the CNN_{stl}. The training losses shown in Figure 14 demonstrate that the framework managed to decrease the CNN loss as well as the overall loss as the training progressed. Additionally, it can be seen from Figure 15a and 15b that the discriminator loss decreases at first and becomes stable indicating that the model learns to discriminate between real and synthetic DM images, and the generator loss seems to increase at first and stabilizes later on when it



(a) Accuracy (b) CNN loss (c) Total loss

Figure 14: Training Losses when training CycleGAN + CNN$_{stl}$



(a) D_A (b) G_A (c) Cyc_A

(d) D_B (e) G_B (f) Cyc_B

Figure 15: Training Losses when training CycleGAN + CNN$_{stl}$ (denoted in orange) and plain CycleGAN (denoted in blue): Smoothing of 0.99 was used in generating curves for losses

starts fooling the discriminator. Experiments on training the modified CycleGAN resulted in an improvement compared to all other approaches.As mentioned in Section 3, a pretrained CycleGAN model (trained with normal patches) was used as a starting point for both the experiments. The reason for such scheme is because we found out that CycleGAN learns the low frequency features that generally describe the two domains early on in the training. Since the number of normal samples was much higher in the training data, a model trained on normal samples was chosen as a starting point. The first experiment was training on the whole training data that contains imbalanced number of abnormal and normal patches. The imbalance was accounted for by assigning a higher weight (λ_c) to the CNN loss for the under-sampled class($\mathcal{L}_{CNN}$). The second experiment was made on a subset of the data with a balanced number of classes (see Table 3). In both experiments, an improvement of 8% was obtained in the validation set. This shows that the network is successfully using the additional information given by the $\mathcal{L}_{CNN}$. Besides giving an external information, adding this loss may have helped the results because the network is less constrained by the cycle-consistency loss. A summary of performance comparison with all the algorithms experimented shown in Figure 11 demonstrates that the modified CycleGAN implementation not only improves the baseline approach, but is somewhat approaching the upper bound performance. This claim is more visible in the testing set results in Figure 13.

Additionally, the performance of UNIT is seen to be the worst both qualitatively and quantitatively. The default training parameters used for this implementation have not helped it to learn to translate regardless of the number of iterations it took. A different set of training parameters could potentially improve the results but more investigation of the working principles might be needed as well. Although the results of modified CycleGAN look promising, one thing to note is that real DM images have overlapping tissues caused as a result of image acquisition techniques as discussed in Section 1. Due to this, the model could learn to introduce syn-

thetic overlaps in the generated DM patches to ensure that they look like the real ones. This could be a disadvantage and more studies might be needed to find an alternative way.

6. Conclusions

This work of research has demonstrated that unpaired image-to-image translation methods can be used in medical imaging, specifically in translating DBT to DM with an acceptable performance. The major findings of this study can be summarized in the following manner. Translation methods based on GANs, such as CycleGAN can be successfully modified to serve a specific task on a given imaging data. To our knowledge, this is the first time image translation models have been applied for domain transferring task between DBT and DM images. Moreover, the suggested modifications on conditioning CycleGAN on data classes have been applied and resulted in acceptable results on the given task. Application of such methods can serve as a means of generating realistic looking synthetic DM images from DBT sources. In such way, existing deep learning based CAD systems trained on DM images can be used to evaluate translated DBT patches. This approach may also serve as an alternative to finetuning with DBT data that has limited availability. Future work should focus on crafting sophisticated evaluation metrics specified on DM/DBT imaging modalities to further assess the characteristics of generated synthetic images and how they could directly compare to unpaired images in the real domain.

7. Acknowledgments

Mahlet Alie Birhanu was supported by the Erasmus Mundus Joint Master Degree program in Medical Imaging and Applications. Acknowledgements are extended to all ScreenPoint Medical staffs who collaborated during this work of research.

References

Arjovsky, M., Chintala, S., Bottou, L., 2017. Wasserstein generative adversarial networks, in: Precup, D., Teh, Y.W. (Eds.), Proceedings of the 34th International Conference on Machine Learning, PMLR, International Convention Centre, Sydney, Australia. pp. 214–223. URL: http://proceedings.mlr.press/v70/arjovsky17a.html.

Armanious, K., Jiang, C., Abdulatif, S., Küstner, T., Gatidis, S., Yang, B., 2019. Unsupervised medical image translation using cycle-medgan. CoRR abs/1903.03374. URL: http://arxiv.org/abs/1903.03374.

Bazzocchi, M., Mazzarella, F., Frate, C.D., Girometti, F., Zuiani, C., 2007. Cad systems for mammography: a real opportunity? a review of the literature. La radiologia medica 112, 329–353.

Becker, A., Jendele, L., Skopek, O., Berger, N., Ghafoor, S., Marcon, M., Konukoglu, E., 2018. Injecting and removing malignant features in mammography with cyclegan: Investigation of an automated adversarial attack using neural networks. CoRR .

Berthelot, D., Schumm, T., Metz, L., 2017. BEGAN: boundary equilibrium generative adversarial networks. CoRR abs/1703.10717. URL: http://arxiv.org/abs/1703.10717.

Broeders, M., Moss, S., Nystrm, L., Njor, S., Jonsson, H., Paap, E., Massat, N., Duffy, S., Lynge, E., Paci, E., 2012. The impact of mammographic screening on breast cancer mortality in europe: A review of observational studies. Journal of Medical Screening 19, 14–25. URL: https://doi.org/10.1258/jms.2012.012078, doi:10.1258/jms.2012.012078. pMID: 22972807.

Chan, H.P., Wei, J., Zhang, Y., Helvie, M.A., Moore, R.H., Sahiner, B., Hadjiiski, L.M., Kopans, D.B., 2008. Computer-aided detection of masses in digital tomosynthesis mammography: comparison of three approaches. Medical physics 35 9, 4087–95.

Chu, C., Zhmoginov, A., Sandler, M., 2017. Cyclegan, a master of steganography. CoRR abs/1712.02950. URL: http://arxiv.org/abs/1712.02950.

Giess, C., Pourjabbar, S., Ip, I., Lacson, R., Alper, E., Khorasani, R., 2017. Comparing diagnostic performance of digital breast tomosynthesis and full-field digital mammography in a hybrid screening environment. American Journal of Roentgenology 209, 1–6. doi:10.2214/AJR.17.17983.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, pp. 2672–2680.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., Hochreiter, S., 2017. Gans trained by a two time-scale update rule converge to a nash equilibrium. CoRR abs/1706.08500. URL: http://arxiv.org/abs/1706.08500.

Isola, P., Zhu, J., Zhou, T., Efros, A.A., 2016. Image-to-image translation with conditional adversarial networks. CoRR abs/1611.07004. URL: http://arxiv.org/abs/1611.07004.

Kooi, T., Litjens, G., van Ginneken, B., Gubern-Mrida, A., Snchez, C.I., Mann, R., den Heeten, A., Karssemeijer, N., 2017. Large scale deep learning for computer aided detection of mammographic lesions. Medical Image Analysis 35, 303 – 312. doi:https://doi.org/10.1016/j.media.2016.07.007.

Korkinof, D., Rijken, T., O'Neill, M., Yearsley, J., Harvey, H., Glocker, B., 2018. High-resolution mammogram synthesis using progressive generative adversarial networks. CoRR abs/1807.03401. URL: http://arxiv.org/abs/1807.03401.

Lauby-Secretan, B., Scoccianti, C., Loomis, D., Benbrahim-Tallaa, L., Bouvard, V., Bianchini, F., Straif, K., International Agency for Research on Cancer Handbook Working Group, de Bock, G., 2015. Breast-cancer screening–viewpoint of the iarc working group. New England Journal of Medicine 372, 2353–8. doi:10.1056/NEJMsr1504363.

Liu, M., Breuel, T., Kautz, J., 2017. Unsupervised image-to-image translation networks. CoRR abs/1703.00848. URL: http://arxiv.org/abs/1703.00848.

Lucic, M., Kurach, K., Michalski, M., Bousquet, O., Gelly, S., 2018. Are gans created equal? a large-scale study, in: Proceedings of the 32Nd International Conference on Neural Information Processing Systems, Curran Associates Inc., USA. pp. 698–707. URL: http://dl.acm.org/citation.cfm?id=3326943.3327008.

M Hakim, C., M Chough, D., Ganott, M., Sumkin, J., L Zuley, M., Gur, D., 2010. Digital breast tomosynthesis in the diagnostic environment: A subjective side-by-side review. AJR. American journal of roentgenology 195, W172–6. doi:10.2214/AJR.09.3244.

Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. CoRR abs/1411.1784. URL: http://arxiv.org/abs/1411.1784.

Odena, A., Olah, C., Shlens, J., 2017. Conditional image synthesis with auxiliary classifier GANs, in: Precup, D., Teh, Y.W. (Eds.), Proceedings of the 34th International Conference on Machine Learning, PMLR, International Convention Centre, Sydney, Australia. pp. 2642–2651. URL: http://proceedings.mlr.press/v70/odena17a.html.

P. Zuckerman, S., Maidment, A., P. Weinstein, S., Mcdonald, E., Conant, E., 2017. Imaging with synthesized 2d mammogra-

phy: Differences, advantages, and pitfalls compared with digital mammography. American Journal of Roentgenology 209, 1–8. doi:10.2214/AJR.16.17476.

Samala, R., Chan, H.P., Hadjiiski, L., A. Helvie, M., Wei, J., Cha, K., 2016. Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography. Medical Physics 43, 6654–6666. doi:10.1118/1.4967345.

Seeböck, P., Romo-Bucheli, D., Waldstein, S.M., Bogunovic, H., Orlando, J.I., Gerendas, B.S., Langs, G., Schmidt-Erfurth, U., 2019. Using cyclegans for effectively reducing image variability across OCT devices and improving retinal fluid segmentation. CoRR abs/1901.08379. URL: http://arxiv.org/abs/1901.08379.

Siegel, R.L., Miller, K.D., Jemal, A., 2018. Cancer statistics, 2018. CA: A Cancer Journal for Clinicians 68, 7–30. URL: https://onlinelibrary.wiley.com/doi/abs/10.3322/caac.21442, doi:10.3322/caac.21442.

Wang, T., Lin, Y., . CycleGAN with Better Cycles. CoRR , 1–8.

Warren Burhenne, L.J., Wood, S.A., D'Orsi, C.J., Feig, S.A., Kopans, D.B., O'Shaughnessy, K.F., Sickles, E.A., Tabar, L., Vyborny, C.J., Castellino, R.A., 2000a. Potential contribution of computer-aided detection to the sensitivity of screening mammography. Radiology 215, 554–562. URL: https://doi.org/10.1148/radiology.215.2.r00ma15554, doi:10.1148/radiology.215.2.r00ma15554. pMID: 10796939.

Warren Burhenne, L.J., Wood, S.A., D'Orsi, C.J., Feig, S.A., Kopans, D.B., O'Shaughnessy, K.F., Sickles, E.A., Tabar, L., Vyborny, C.J., Castellino, R.A., 2000b. Potential contribution of computer-aided detection to the sensitivity of screening mammography. Radiology 215, 554–562. URL: https://doi.org/10.1148/radiology.215.2.r00ma15554, doi:10.1148/radiology.215.2.r00ma15554. pMID: 10796939.

Wolterink, J.M., Dinkla, A.M., Savenije, M.H.F., Seevinck, P.R., van den Berg, C.A.T., Išgum, I., 2017. Deep mr to ct synthesis using unpaired data, in: Tsaftaris, S.A., Gooya, A., Frangi, A.F., Prince, J.L. (Eds.), Simulation and Synthesis in Medical Imaging, Springer International Publishing, Cham. pp. 14–23.

Zhu, J., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. CoRR abs/1703.10593. URL: http://arxiv.org/abs/1703.10593.

Medical Imaging and Applications

Master Thesis, June 2019

Patient Motion Correction in Digital Subtraction Angiography

Brianna Burton, Eva Vandersmissen

Agfa NV, Septestraat 27, 2640 Mortsel, Belgium

Abstract

Digital subtraction angiography (DSA) is a technique for the visualization of blood vessels in the human body using a contrast agent. X-ray images, obtained before and after the injection of a contrast agent into the vessels, are subtracted to remove anatomical structures in the images. The resultant images show only the contrast filled blood vessels without any background structures. However, this technique is very sensitive to patient motion. Any movement will become visible as artifacts in the final subtraction images, reducing their diagnostic value. To improve the quality of subtracted images, various registration methods have been proposed. In this thesis, several approaches to image registration for digital angiographic images are investigated. A new fast registration algorithm based on a discrete Fourier transform is proposed to accurately generate artifact free subtraction images. The resultant subtraction images are then used to train a neural network to compare the applicability of deep learning for virtual subtraction with the proposed registration method. Experimental results with a clinical dataset show the improvement in image quality with image registration as well the potential of virtual subtraction.

Keywords: x-ray angiography, image registration, deep learning, cross correlation

1. Introduction

In interventional radiology, angiography is an x-ray technique employed for the visualization of blood vessels in a bony or dense soft tissue environment. Contrasting with traditional x-ray projection images, which provide little to no contrast between the surrounding tissue and vessels, angiography enhances visibility using a radiopaque contrast agent that is injected into the target vessels, and the image obtained includes the blood vessels and all of the the under and overlying structures. It is utilized to identify vascular abnormalities, detect injury in vessels after trauma, or evaluate the vasculature of a tumor before surgery and is essential to a physician's diagnosis (Lee et al., 2019). However, angiographic images still suffer from low contrast between the vessels and the surrounding tissue.

To remove the distracting structures, digital subtraction angiography (DSA) can be used. An iodine based contrast agent is injected into the patient's blood vessels. This agent appears highly opaque in x-ray images due to its high x-ray absorption rate. Concurrently, a sequence of images is taken to show the inflow of contrast into the vessels of interest. The sequence begins by obtaining a *mask* image, absent of contrast, and subsequent *live* images, as the contrast is injected. The mask and live images only differ by the opacified blood vessels. Therefore, the mask image can be subtracted pixel by pixel from each live image, removing anatomical structures and leaving behind only the blood vessels filled with contrast, as shown in Fig. 1.

Since its commercialization in 1980, DSA has improved diagnosis and treatment for various arterial and venous occlusions (Crummy et al., 2018). Open vascular procedures have been substituted by minimally invasive endovascular procedures, with acquisitions that can be viewed immediately. Such applications include endovascular anuresym repair, arterial balloon angioplasty, arterial stenting, endovasuclar embolization, and thrombectomy. DSA has high temporal and spatial resolution, which is unmatched in other imaging modalities (Crummy et al., 2018).

Despite the increase in popularity of 3D imaging techniques with virtual angiography, DSA remains the gold standard for diagnosis of aneurysms because of its ability to detect the micro aneurysms missed by other

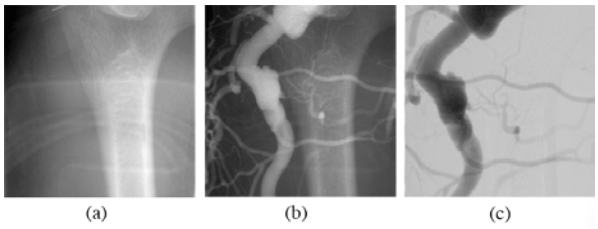


Figure 1: Example of images in a DSA pipeline: (a) is the mask image, (b) the live image, and (c) the subtraction of the two, where the vessels are clearly visible.

modalities (Wang et al., 2015). It is also preferred in some cases to computed tomography (CT) or magnetic resonance (MR) images, as it can better diminish the effects of the skull and hematoma in images. A recent study comparing DSA to MR and CT angiography found that DSA resulted in an overall better outcome for the patient and were more cost-effective over time (Sailer et al., 2013). DSA can also be used in the operating theater, while a patient is undergoing surgery, unlike 3D techniques.

Although commonly used clinically, the subtraction technique in DSA assumes that there is no change in the background anatomical structures during acquisition. Any patient movement, including instances like respiration, swallowing, cardiac motion, or intestinal gases, will become visible as misalignment artifacts in the subtracted images. Clinical evaluations have proved that patient motion always occurs (Meijering et al., 1999). Distortions due to movement limit the quality and amount of diagnostic information that can be extracted from the images. Since the 1980s, techniques such as dual energy subtraction, and automatic masking have been developed to attempt to improve diagnostic value by removing artifacts, but require special devices and have never been introduced on a large scale (Meijering et al., 1999). Various image processing techniques have also attempted to retroactively correct for these artifacts before subtraction (Meijering et al., 1999). However, to date, proposed algorithms do not efficiently compensate for complex patient motion, limiting their clinical applications.

Despite various literature on image registration for DSA, sufficiently fast and accurate methods for integration into clinical applications have not been proposed. In this thesis, several methods to register angiographic images are explored, and a new registration method is proposed for fast clinical applications. Then, motion corrected images are used to train a convolutional neural network and virtually subtracted images are generated, eliminating the need for image registration. In Section 2, the state of the art for patient motion correction in DSA is described. Section 3 describes the proposed method. Results of experiments on angiographic image sequences, with the proposed and convolutional neural network approach are presented in Section 4 and discussed in Section 5. Conclusions are presented in Section 6.

2. State of the art

2.1. Analysis of Patient Motion

Patient motion that manifests as artifacts in DSA images is rarely uniform: for example, in an image including the heart and lungs, different parts of the image may be influenced more by the movement of one or the other. Typically, artifacts only appear in regions where strong object edges are present, due to misalignment between the mask and live images (Meijering et al., 1999). To compensate for this motion, a *pixel shift*, or shift of the entire image in the x and/or y directions may be able to remove distortions in one part of an image, but not completely eradicate them throughout. Instead, pixel shifting can be applied in a region of interest (ROI), where the movement is smaller, and the background structures may be removed.

To approximate patient motion in an entire image, it is thought that non-rigid local transformations can better approximate local motion. However, due to the nature of x-ray images, which are two dimensional projection images of a three dimensional scene, it has been believed that registration techniques that try to recover a correspondence between two projection images are unlikely to succeed. In a proof by Fitzpatrick (1988) and explained by Meijering et al. (1999), it was shown that because projection images are created by the intensity of x-rays incident on the detector, there exists a two dimensional transformation that completely describes the changes in the images caused by three dimensional motion. Therefore, a 2D vector field can represent motion caused by the patient in a 3D scene. There are several limitations to this theorem, though: discrete image data, the aperture problem (inability to find the tangential velocity), and the fact that new particles (the contrast) are introduced to the scene (Meijering et al., 1999). Therefore, although a transformation exists, these issues limit any kind of registration algorithm for DSA images.

2.2. Clinical Settings

Patient motion in subtracted images is a common issue today in clinical scenarios. Therefore, in current x-ray diagnostic imaging systems, there is often a feature for improving image quality. Most systems have functionality for a simple pixel shift, which addresses rigid motion by shifting the entire image manually or automatically in the x and y directions, or a function that can register small regions of interest non-rigidly (Lee et al., 2019). Clinical methods also must be performed in real-time. In most clinical systems, it required that a user choose a rectangular zone to serve as an ROI, and the registration is limited to that region, accounting for

only a small region's motion. Some examples of algorithms in clinical systems will be described below.

Philips' patent US4870692 includes a method that corrects motion in subregions in an image through cross-correlation. A shift vector is found from the mask sub-image to the live sub-image by means of cross-correlation and the mask sub-image is shifted accordingly. In this method, it has been found that shift vectors are not always reliable and therefore, they can be replaced by neighboring, more reliable shift vectors.

GE holds patent US10985803 for a patient motion correction algorithm using landmarks defined with regular geometrical patterns in the mask image. A registration is performed, and depending on the robustness of the landmarks, the geometric region may be subdivided and registered again.

Additionally, Siemens' patent US8299413B2 describes an automatic calculation of pixel shift vectors, where potential shift vectors are ranked based upon the sum of squared differences metric between the original subtracted image and subtracted images at different shifts, and the best one is applied to the image.

Pixel shifting solutions have been widely adopted in x-ray imaging systems, however they only address rigid motion through pixel shifts. In the literature, though, many solutions to handle non-rigid motion have been described.

2.3. Classical Motion Correction Techniques

Motion correction techniques can be classified into two major categories: *extrinsic* (relying upon artificial objects or markers added to the scene) or *intrinsic* (relying upon anatomical structures) (Markelj et al., 2012). Since the motion in DSA imaging results from anatomical structure movement, intrinsic registration is required. Intrinsic registration can be divided into three main types: feature, intensity, and gradient based methods, and applications of these registration methods to DSA will be described below.

2.3.1. Feature Based Registration

Feature based registration methods minimize the distance between features (point sets) in two images (Markelj et al., 2012). These methods rely heavily upon the quality of anatomical landmarks which make up point sets. In DSA imaging, point sets are unknown and groundtruth data is not available, making this sort of registration unsuitable.

2.3.2. Intensity Based Registration

Intensity based registration techniques rely on only the voxel intensity information in images. In contrast to feature based methods, the similarity measures used are calculated using pixel-wise comparisons (Markelj et al., 2012). For example, Meijering et al. (1999) introduced an intensity based method to compensate for the non-rigid motion, which has since become the basis for many registration algorithms. The correction is performed with two operations: first correspondences are calculated between pixels in the mask and live images, and secondly, a correction is performed by warping one image with respect to the other. Correspondences are defined as a set of *control points* in the mask image and their matches in the live images. Control points are found in the mask image to avoid issues related to vessels present in the live images. Since DSA image artifacts often appear around strong edges, image edge information (gradient magnitude maxima) is used to determine the locations of control points in the mask image. Additionally, criteria is set for a minimum distance between control points.

Next, the correspondences must be found in the live images. For each live image in the exam, correspondences are calculated independently. To match correspondences, a *template matching* approach is used. As opposed to other popular methods such as optical flow, template matching is not impacted significantly by the inflow of contrast into the image (Meijering et al., 1999). Template matching is based on the assumption that for a pixel in one image, its correspondence in the second image can be approximated by searching in a neighborhood and optimizing a similarity measure. Here, the similarity criteria chosen was the histogram of differences, which is robust against the influx of contrast and the window size chosen was 51×51. Once all the correspondence have been found, linear interpolation is used to use the correspondences to warp the final image through a Delaunay triangulation.

Nejati et al. (2013) have also built upon this technique with multi-level b-spline interpolation. Instead of using a gradient magnitude maxima image to find correspondences, a Gaussian derivative image in combination with a Harris corner dector is used. Optimized template matching via a hill climbing algorithm is used to find match points. A multi-level b-spline interpolation is used to warp the mask image to the live image.

All of the previously mentioned methods incorporate template matching, which even when optimized, is still an extremely time consuming operation, when considering that checking just 10% of matches in a 50×50 window still involves 250 calculations, and it is still possible that the best match was missed, unless an exhaustive search was performed.

Liu et al. (2018) performed coarse registration with SURF features for correspondences, and RANSAC to eliminate wrong matches.

However, most of these methods assume unique correspondences, and they may fail to find correspondences when newly visible vessels appear. Other methods try to select features independent of the vessels. Several methods based on 3D space time detection were implemented to avoid features being placed on vessels (Bentoutou and Taleb, 2005; Zhang et al., 2010). Con-

Table 1: Reported computation times for the reviewed methods for images sized 512×512

Method	Time (s)	Notes
Meijering et al. 1999	8.4	
Nejati et al. 2013	14.2	
Nejati and Pourghassem 2014	314	
Liu et al. 2018	0.25	Coarse registration
Lee et al. 2019	1.3	3.30Ghz Core i5 CPU

trol points are selected using image edges, and their intensity values are recorded over time to determine if a feature point's intensity changes due to patient motion or contrast flow. With a 3D space time method, though, the method of rejection for control points depends on the entire exam sequence, and all of the data must be gathered before registration occurs. It also cannot be performed on a subset of images from a sequence.

Multiresolution techniques have also been applied for intensity based registration. For example, Nejati and Pourghassem (2014) performed registration with intensity variation modeling in 4 levels with various sub-blocks. Multiresolution methods, however, suffer due to their long computation time.

2.3.3. Gradient Based Registration

Gradient based methods are based upon the assumption that important image information is contained in the gradients. Since in DSA images the artifacts are often found where edges are present, registration based on gradient images is thought to be suitable. In a paper by Hiroshima et al. (2001), the Laplacian filter is applied to both the mask and live images. Then, a distortion vector is found detecting the peak position of cross-correlation. Many distortion vectors are found on a regular grid, and then b-spline interpolation is used to warp the final image.

2.3.4. Other Registration Methods

Similar to the previously mentioned gradient based method, Lee et al. (2019) registered images by finding the phase difference in the frequency domain without using the gradient. It is assumed that regions in common in two images where motion has occurred will have similar magnitudes but shifted phase. Regions where newly visible vessels appear will have different magnitudes. Therefore, both the mask and live images are transformed into the frequency domain, the magnitudes are matched, and the phase differences are adjusted.

As computational time is of great importance for clinical DSA systems, a comparison of the above described approaches in terms of time is given in Tab. 1.

2.4. Deep Learning Techniques

Since in DSA, the desired result is not a registered image itself, but the subtraction image, it is thought that avoiding registration is possible. For example, virtual DSA can be performed by estimating the subtracted image from the live image. For these methods to succeed, the vessels must be segmented and then the background estimated in those regions (inpainting). Unberath et al. (2017) describe a method using a convolutional neural network (CNN) and U-net architecture to perform image inpainting for background estimation. First, the vessels are segmented using Hessian based segmentation. Then, the neural network is used to perform inpainting in the small region defined by the segmentation mask. In this paper, only simulated images were used to train a network for inpainting. Additionally, the images themselves had smooth backgrounds with little intensity variation and structures. As our dataset included many cases in extremities with bones as prominent features, it can be inferred that the proposed method image inpainting would not perform well on our data.

Another method employed U-net for generating DSA-like images in real time directly from the contrast enhanced live images (Eulig et al., 2019). The CNN was trained with patches of the live and subtracted images. This method obtained conventional DSA level results, but relies on the assumption that the subtracted images used for training are artifact free. Correspondence with the author confirmed that the data itself was not of perfect quality, but images with patient motion were excluded in the study. Given a large number of subtracted images of minimal patient movement, it may be possible to train a neural network to generate its own DSA images.

3. Material and Methods

There are two main goals of the proposed thesis work: the first is to determine the most applicable classical motion correction technique for fast registration of full size DSA images. The second is to use motion corrected images to train a CNN for virtual subtraction. Finally, the results of the two main goals can be compared. The entire framework for the work was implemented in Python. All methods excluding the deep learning framework were executed on a 2.6 GHz Core i5 CPU.

3.1. Data Acquisition

High resolution imaging data was acquired for use in this study from a hospital [name withheld] between 2012 and 2019. Images were acquired from a multimodal x-ray machine similar to the DR800 (Agfa, Belgium). A contrast agent with iodine concentration in the range of 150-300 g/mL was administered to the patient as the sequence was acquired. A total of 9 patients were included in this study. Each patient had between 1 and 24 exams, each exam consisting of between 10-50 images. In total for the 9 patients, there were 79 total exams with 3,312 images. The exams were performed on various body parts including interventions in

the abdomen as well as peripheral exams of the the head, legs, and knees, creating a diverse data set, with assorted amounts of patient motion.

The images were stored in RAW format to preserve patient anonymity and data integrity. The images were 16 bit and ranged in size from 512 × 512 to 1512 × 1512 with voxel resolution 0.25 mm. All but one image were in the linear intensity domain, and this image was examined separately. Because some of the data was acquired during the thesis work, 13 exams from 3 patients were used for the first goal of motion correction. The rest of the data was later included in the final algorithm validation and deep learning approach.

As this dataset of DSA images had not been previously curated, it required full analysis. Since the first and last few frames of a DSA sequence can be noisy or have a skewed intensity range, it was necessary to remove those images from the full set. Additionally, a few patients also had fluroscopy exams, which needed to be separated from the angiography data. Lastly, since the data was in a RAW format, the collimation windows also had to be manually segmented. These images were cropped on an exam basis - all images in an exam have been cropped to the same dimensions.

3.2. Pixel Shift

Since conventional x-ray imaging systems are equipped with a global pixel shifting mode, it was chosen to recreate this system to observe whether motion could be corrected by translation only. The developed system is interactive and by using the arrow keys, the mask image can be shifted in the x and y directions pixel by pixel.

3.3. Elastix

As with many registration problems, it is often useful to explore currently available tools. Elastix is a versatile open source software tool for medical image registration and was chosen to visualize the type of patient motion in DSA sequences. The goal was to perform the best registration possible, regardless of computation time. It was chosen to perform registration with an affine, then b-splines registration. The affine registration provides a rough first alignment. As b-splines is a non-rigid registration technique, it is possible to recover some of the non-rigid movements in the images. Parameter files for registration were implemented from the Elastix database (http://elastix.bigr.nl), with mutual information and advanced normalized correlation as similarity metrics.

3.4. SIFT

SIFT is another popular algorithm for image registration as the features it generates are scale and rotation invariant, and Liu et al. (2018) showed that registration for DSA images is possible with a similar algorithm, SURF.

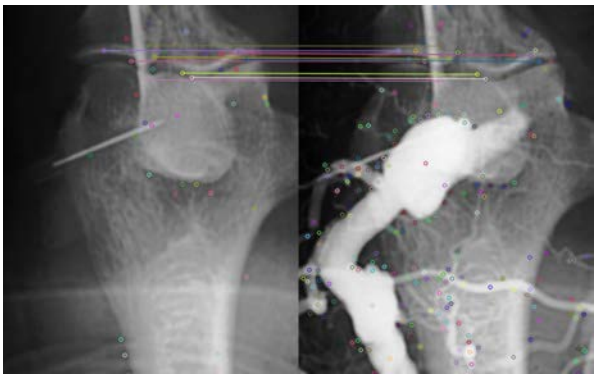


Figure 2: SIFT keypoints and matches in the mask and live images.

Therefore, it was chosen to implement SIFT-based registration with OpenCV. Since SIFT and SURF features are both patented, it was chosen to also test open source ORB features as well. The images were first normalized to zero mean and unit variance. More preprocessing was required to obtain any keypoints with SIFT or ORB. The preprocessing step consisted of anisotropic diffusion, histogram equalization, and image sharpening. The threshold for ORB features was increased to 100000 to obtain as many features as possible. Example SIFT features and their matches can be seen in Fig. 2.

3.5. Smart Control Points

As a significant portion of the literature published about DSA registration incorporates feature-based control point selection along with template matching, it was essential to implement this type of method. Of the reviewed literature, the method defined by Nejati et al. (2013) is the most recent and complete explanation of this task and was implemented by the author. An overview of the algorithm will be provided below.

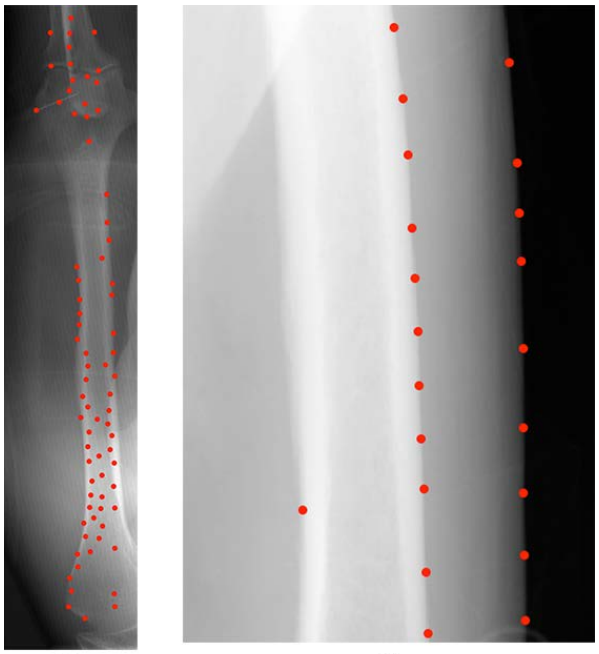
The mask image is used for control point detection as it contains no contrast and the same mask is used for each live image. First, the gradient magnitude of the mask image was detected using the derivative of the Gaussian at scale $\sigma = 1$ and normalized in the range [0,1]. Then, the strongest edges were found by thresholding the image at the mean, obtaining a final image, which will be referred to as G. Then, the Harris corner response is calculated as:

$$R = \hat{I}_x^2 \hat{I}_y^2 - \hat{I}_{xy}^2 - k(\hat{I}_x^2 + \hat{I}_y^2)^2 \tag{1}$$

where

$$\begin{aligned} \hat{I}_x &= I_x * g \\ \hat{I}_y &= I_y * g \\ \hat{I}_{xy} &= (I_x \dot{I}_y) * g \end{aligned} \tag{2}$$

where I_x and I_y are partial derivatives of an image I in x and y, g is a Gaussian window, and $*$ is a convolution. The sensitivity parameter k is set to 0.12. All negative values of R are set to 0 and then normalized in the interval [0,1]. A weighted average of R and G is calculated



(a) (b)

Figure 3: Smart control point selection in two mask images. (a) shows smart control points that cluster in a bright region and (b) shows smart control points that appear only on one edge of a bone.

as

$$\hat{R} = \alpha G + \beta R \tag{3}$$

where $\alpha = 0.3$ and $\beta = 0.7$. The local maxima greater than $t_1 = 0.1$ of $\hat{R}$ are found in neighborhoods of radius 5. Then, the list is sorted from largest to smallest. The list is filtered by removing all points in a radius of 45 pixels around the highest control point and moving through the list, so only the largest points with a minimum distance 45 remain. These points are the control points. An example of control point selection can be seen in Fig. 3

To find match points in the live image, template matching is used. For each control point, a 50×50 window is centered on the control point in the mask image and a metric is used to compare to a 50×50 window in the live image. A $50\times$ 50 window is optimal as it leads to smooth match surfaces (Nejati et al., 2013). The window is shifted with a maximum distance 20 pixels in each the x and y directions, as patient motion is subtle. The center pixel over which the optimized metric is found represents the match point in the live image. Mutual information and the entropy of the normalized histogram of differences were both tested for the template matching algorithm. Therefore, each control point has a corresponding matching point in the live image.

3.6. Discrete Fourier Transform

Phase correlation is another popular technique to correct gross translations in image registration. Hiroshima et al. (2001) showed that an extension of this method based on local distortion vectors works well for head, neck, abdominal, and leg images. Distortion vectors are calculated through phase correlation in many ROIs in an image. Therefore, this method can account for local deformities, like rotation, contraction, and relaxation. As the deformation model is not as sophisticated as smart control point methods, it can be calculated more efficiently. This method has been adapted by the author, and an explanation of the implemented method will be described below.

The mask and live images are preprocessed to enhance image edges, further details are given in Section 3.6.1. An ROI of size $n \times n$ is centered at position $[i, j]$ in the mask and live images. The discrete 2D Fourier transform of each ROI is found. The normalized cross power spectrum of these Fourier transforms is calculated as

$$R = \frac{\mathbf{G}_l \circ \mathbf{G}_m^*}{|\mathbf{G}_l \circ \mathbf{G}_m^*|} \tag{4}$$

where $\mathbf{G_l}$ and $\mathbf{G_m}$ are the Fourier transforms of the live and mask images respectively, $*$ is a complex conjugate, and $\circ$ is entry wise multiplication (Foroosh et al., 2002). Normalized cross-correlation is found by finding the inverse Fourier transform of R, obtaining r. The peak of r in x and y is one distortion vector from the mask to the live image. Subpixel registration can be obtained by performing interpolation to non-integer values for the peak location.

Once a distortion vector is found, the ROI is shifted by M in both the x and y directions, and the process is repeated, until $m \times n$ (image size dependent) distortion vectors are found. The center points of all ROIs can be considered the control points and by adding each distortion vector to its control point, match points are found in the live image.

A window size of 151×151 was implemented by Hiroshima et al. (2001), and the author has chosen sizes 25, 50, 100, 150, 200, and 300 for comparison. The window spacing, M, selected must be smaller that the corresponding window size. Window spacings of 75, 100, 150, and 200 were compared.

3.6.1. Preprocessing

As x-ray images are noisy and image edges are known to be the location of many artifacts in DSA, preprocessing to enhance edges is essential. As a first step, all images in an exam were normalized by subtracting the mean intensity and dividing by the standard deviation of the mask image. Gaussian smoothing and anisotropic diffusion were compared. Sharpening and histogram equalization were also added. Several edge detection methods were tested including Laplacian, Sobel, Gaussian derivative ($\sigma = 1$), and Canny edge filters. Both signed and unsigned versions of Canny edges were compared. Signed canny edges were found in the horizontal direction (artifacts are mostly found along long

vertical edges), depending on an increase or decrease in intensity along that edge. Temporal averaging of pairs of images was also used to reduce noise, which limits the amount of images in a DSA series by half.

From preliminary testing, it was found that Gaussian smoothing, anisotropic diffusion, and sharpening did not affect DFT registration, and the edge detection methods with the most potential were the Gaussian derivative and Canny edge detection (with high threshold). Fig. 4 shows the preprocessed images before registration. A typical preprocessing pipeline would contain normalization and edge detection. The final edge image underwent the discrete Fourier transform.

3.6.2. Smart Control Points

In addition to template matching, smart control points as found in the mask image were also matched to the live image using DFT. The method to find distortion vectors is the same as described in Section 3.6, but instead of windows centered on a regular interval, the window of 200×200 are centered upon the smart control points and they are matched to the live image using a DFT.

3.7. Image Warping

For DSA registration, it is essential to warp the mask image to each live image, and not the reverse, as there is a flow of contrast into the scene and the contrast should remain unchanged. In the case of smart control points and the discrete Fourier transform, there are two point sets: one in the mask image and one in the live image. It is necessary to have a dense correspondence between the mask and live images which can be estimated through the interpolation of the control points to warp the mask image to the live image. Here, a thin plate splines (TPS) interpolation function was used because TPS is optimal for interpolating an image when given a set of points at irregular intervals. A complete description of TPS warping is given by Nejati and Pourghassem (2014).

3.8. Evaluation Metrics

In many medical imaging problems, quantitative evaluation is essential. To calculate quantitative similarity when registering images, correspondences can be calculated between two images, when the true correspondences are known. Standard quantitative measures to assess the quality of DSA images do not exist (Lee et al., 2019). As with DSA images there is no groundtruth data, alternative quantification methods must be used. In a study by Nejati et al. (2013), simulated mask images were generated from live images with little movement through vessel segmentation, and a known geometric transformation was applied to the live image. Coordinates in the mask image were selected, random shifts of $+/-10$ pixels were added, and the live image was warped according to these shifts using b-spline interpolation. The shifts were recovered using template matching and a similarity metric. The similarity measures entropy of the histogram of differences (EHD), cross-correlation (CC), mutual information (MI), mean structural similarity (SSIM), and entropy of the normalized histogram of differences (ENT) were compared. The root mean squared (RMS) error was compared between the known original coordinate positions and those found by template matching. It was found that ENT resulted in the lowest RMS error, indicating that it is a good similarity measure for DSA images.

Much of the quality of registration is evaluated visually (whether there are artifacts or not). When comparing a registration methods, is possible to evaluate the difference between two by comparing the difference in a metric before and after registration.

3.9. Virtual Subtraction

Virtual subtraction was performed using a U-net architecture and Keras functional API with Tensorflow as a backend. The architecture combines downsampling and a high resolution output via skip-ahead connections, ideal for the small vessels in DSA images (Unberath et al., 2017). The U-net consists of an encoder portion, where downsampling layers decrease the feature size as the depth increases. Each layer has 2 convolutional layers (kernel size 3×3, stride 1×1, padding 1×1) with a rectified linear unit (ReLU) as nonlinearity followed by a maxpooling layer (2×2). In the maxpooling layer, the size of the features is decreased by a factor of 2, but in the convolutional layer, the number of features is doubled. Therefore, as the depth increases, the features lose spatial and gain contextual information. The U-net also has a decoder that transmits contextual information through upsampling layers. Each upsampling layer halves the number of features while the number of dimensions doubles, and skip-ahead connections combine the output of upsampling with features from the downsampling layer to gain contextual information. Therefore, the decoder mirrors the encoder in structure. A total of 6 layers were used for all experiments and the architecture is shown in Fig. 5. Finally, a convolution is performed for each individual pixel to output the final mask in greyscale.

3.9.1. Data Preparation

Images were divided into 224×224 non-overlapping patches. The mask image was registered and subtracted from the live image using cross-correlation and pixel shifting, and some training patch examples are found in Fig. 6. The optimal preprocessing was selected and applied before calculating the shift. Subtracted images were normalized to the range [0,1] with the background set to 0.7 to ensure consistency. The training data consisted of 22,136 image patches.

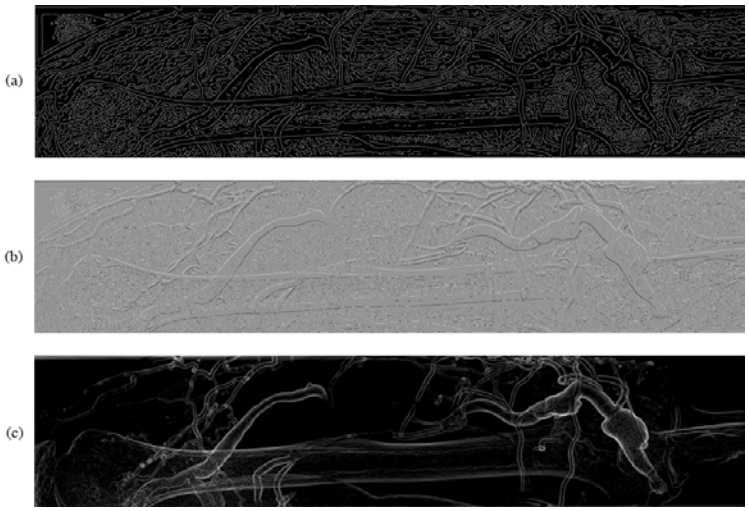


Figure 4: Preprocessing using (a) canny unsigned edge detection, (b) canny signed edge detection, and (c) Gaussian derivative.

3.9.2. Training

The data was divided in training, validation, and test sets by patient to avoid any overlap in data between the sets. The U-net was trained using the Adam optimizer with both L1 and L2 loss being compared. All weights were initialized using He initialization, and the data was augmented using flips, rotations, shears, and scaling. The network was optimized by changing the learning rate, adding dropout, and modifying the amount of training. The training pipeline was implemented on an NVIDIA Tesla V100 DGXS.

4. Results

4.1. Elastix

Elastix registration provided an insight to the type of patient movement common in DSA images. In only one case, the registration was successful and resulted in nearly complete artifact removal as seen in Fig. 7. This figure also illustrates the non-rigid registration required to compensate for the movement. For most other cases, however, the results were not consistent. Elastix registration enhanced artifacts in images that were previously not visible in the subtracted image. Additionally, registration of one image took approximately 30s.

4.2. SIFT and ORB

In SIFT and ORB feature detection, it was determined that preprocessing images with sharpening was most important to find keypoints. When comparing the SIFT and ORB feature detectors on the same images, SIFT found significantly more keypoints than ORB. However, neither feature detector was able to detect a large amount of keypoints in images with long bones. Instead, keypoints were often detected along other insignificant points of the image (where there was no initial movement). The most common locations of keypoint detection were along the contours of the body. As the movement usually manifests along bony structures, it was determined that this method cannot be applied to DSA images.

4.3. Smart Control Points

Unlike SIFT or ORB features, smart control points were detected where significant movements were found in an image (along the long, bony structures). However, when template matching was performed to find match points with both MI and ENT, the matches found were often incorrect, and the images were not able to be registered. This was confirmed on every tested case.

4.4. Discrete Fourier Transform

The metrics of various DFT experiments can be found in Tab. 2. Comparisons of the methods as shown in one image can be found in Fig. 8. By comparing the metrics as well as the images, it was concluded that the best performance was obtained using Canny signed preprocessing, window size of 200×200, and window spacing of 150. For 512×512 images, the average computation time was 0.6927s, and over the full dataset, the computation time ranged from 0.5468s to 8.5156s depending on image size. Examples of this registration applied to images of various types can be seen in Fig. 9.

4.5. Virtual Subtraction

Fig. 10 shows the application of virtual subtraction to various image patches, compared with the groundtruth (registered image), and subtracted image before registration.

5. Discussion

5.1. Classical Motion Correction Techniques

Overall, it can be stated that patient motion correction in a variety of DSA images has been achieved. From

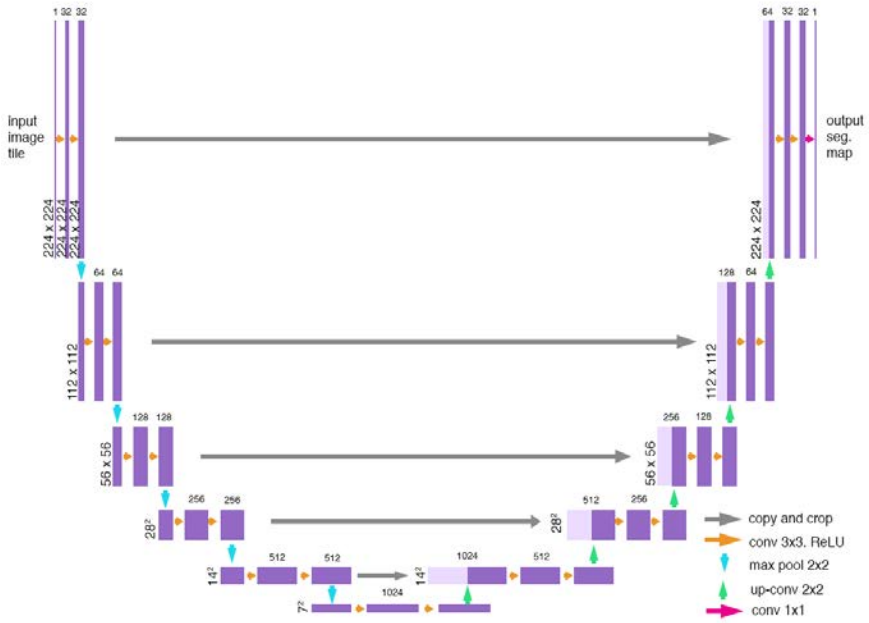


Figure 5: U-net architecture employed here.

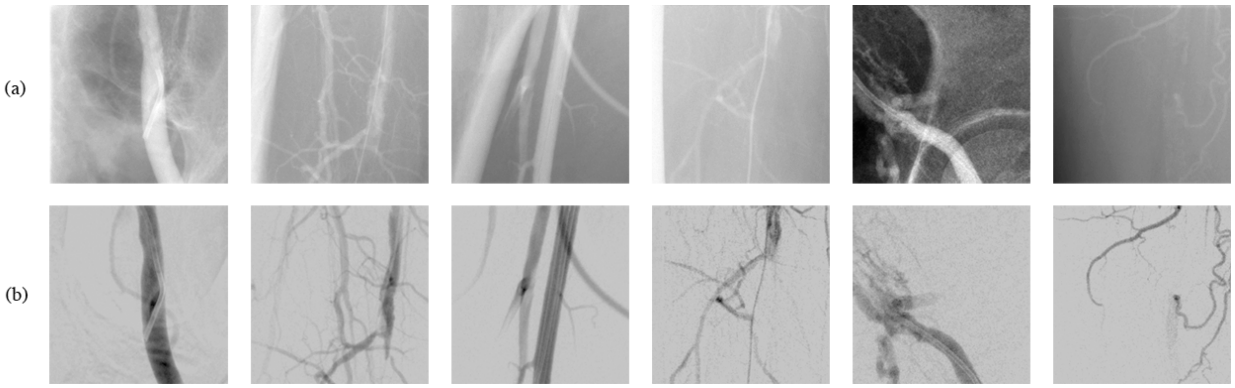


Figure 6: Example image patches used to train U-net. (a) shows the live image patch and (b) shows the corresponding subtracted image patch.

the first experiments with Elastix data, as seen in Fig. 7, it can be seen from the deformation field that typical patient motion is not global, and a non-rigid registration method would be best for correcting this motion. The deformation field also shows that the movements are extremely small. Elastix had trouble distinguishing which regions in an image should be registered, perhaps due to the metric of mutual information. Since contrast is introduced between the mask and live images, it is possible that MI is not an appropriate metric, since it looks for correlation between pixels, which would not be present for the contrast areas. When another metric, AdvancedMeanSquares, was tested, it did not encounter the same issues. Elastix software was not practical due to the long (30s) registration time.

When considering registration using SIFT or ORB features, it is seen that very few keypoints were found. Therefore, this method suffered during registration, since there were no control points in regions that required registration. This may be because SIFT keypoint detection focuses on corners, and specifically discards keypoints along edges, whereas with DSA images, often the motion is found only along those long edges. This method was unable to compensate for any types of motion.

From Fig. 8, it is easy to see that testing various proposed methods for the preprocessing and control point selection was essential. For preprocessing, it can be seen that with Canny preprocessing, which produces a binary image, the addition of a sign based on the edge direction has helped prevent mis-registration. This is likely due to the extra edge information, giving significantly more data for registration to occur. When comparing Canny signed and Gaussian derivative preprocessing, from Tab. 2, it can be seen that the ENT metric is much less for the Canny signed preprocess-

Table 2: Comparing different methods for DFT based registration

	Preprocessing	Window Size (px)	Window Spacing (px)	ENT Difference	MI Difference	Average Time (s)
Changing Preprocessing Methods						
AD	Canny	200	200	-0.1287	0.0140	
AD	Canny Signed	200	200	-0.1296	0.0121	
AD	Gaussian Derivative	200	200	-0.0821	0.0077	
No AD	Canny	200	200	-0.1197	0.0120	
No AD	Canny Signed	200	200	-0.1286	0.0110	
No AD	Gaussian Derivative	200	200	-0.0908	0.0001	
Changing Window Size and Spacing						
	Canny Signed	100	100	-0.1514	-0.0141	
	Canny Signed	150	150	-0.1188	0.0027	
	Canny Signed	200	200	-0.1286	0.0110	
	Canny Signed	300	300	-0.0986	0.0050	
	Canny Signed	150	75	-0.1400	0.0067	
	Canny Signed	150	100	-0.1400	0.0076	
	Canny Signed	200	75	-0.1339	0.0086	
	Canny Signed	200	100	-0.1300	0.0143	
	Canny Signed	200	150	-0.1076	0.0098	
Adding Temporal Averaging						
No TA	Canny Signed	200	150	-0.0966	0.0115	
TA	Canny Signed	200	150	-0.1234	0.0200	
Comparing with Smart CPs						
Smart CPs	Canny Signed	200	200	-0.0336	-0.1261	6.6716
Grid Based	Canny Signed	200	200	-0.1286	0.0110	2.7307

AD - Anisotropic diffusion, TA - Temporal averaging

ing method. Combined with visual results from the full dataset, Canny signed preprocessing was chosen as the best preprocessing technique.

In comparing the window size, from Fig. 8 it is seen that larger window sizes have more consistent registration, and less artifacts are created during registration. However, there is a limit to this, as in Fig. 8.h (window size 300×300, where the long bone cannot be removed due to the large window size. Additionally, although Tab. 2 indicates the best performance is with a 100×100, Fig. 8.g shows that this size can create artifacts. As a consistent metric has not been established for DFT registration, the metrics presented in the table cannot be taken as absolute indications without also referring to the visual results. From the experiments in window spacing and size, it can be determined that window size 200×200 and window spacing 150 indicate the most consistent performance. When comparing the proposed method to a method presented by Lee et al. (2019), the computation time for the same size image is nearly twice as fast for our method, while achieving similar visual results using comparable architecture.

The addition of temporal averaging did indicate an increase in performance, but significantly reduced the amount of data, and as DSA images are high resolution with small vessels, it is thought that temporal averaging may cause some small vessels to become less contrasted with the surrounding tissue when averaged. Therefore, this method could not be used in clinical settings.

Fig. 8.f shows the application of DFT using the smart control points, rather than a regular grid. According to Tab. 2, the ENT metric has a larger decrease using the grid based method rather than smart control points. When one looks at the smart control points in Fig. 8.e, it is seen that one edge of the bone was not selected, and a smart control point is even found on the catheter, which has no movement (since it is not seen in the original subtracted image). Although there are so many on one side of the bone, they do not produce any registration there, thus the points do not accurately represent the image. This may be due to the weighting of the Harris corner response in Equ.3, as more weight is given to the Harris corner response. Since artifacts are observed on image edges (bones), the Harris corner response may not be as effective at finding control points on straight edges, since it is a corner detector. By considering both the metrics and visual results, smart control points are not as consistent as a grid based method.

From the above, it can be seen that the grid based method had advantages over the smart control points method. Since larger window sizes indicate more consistent registration, a window size of 200×200 was chosen, with a window spacing of 150, to be applied to all images. Fig. 9 shows that these settings could register

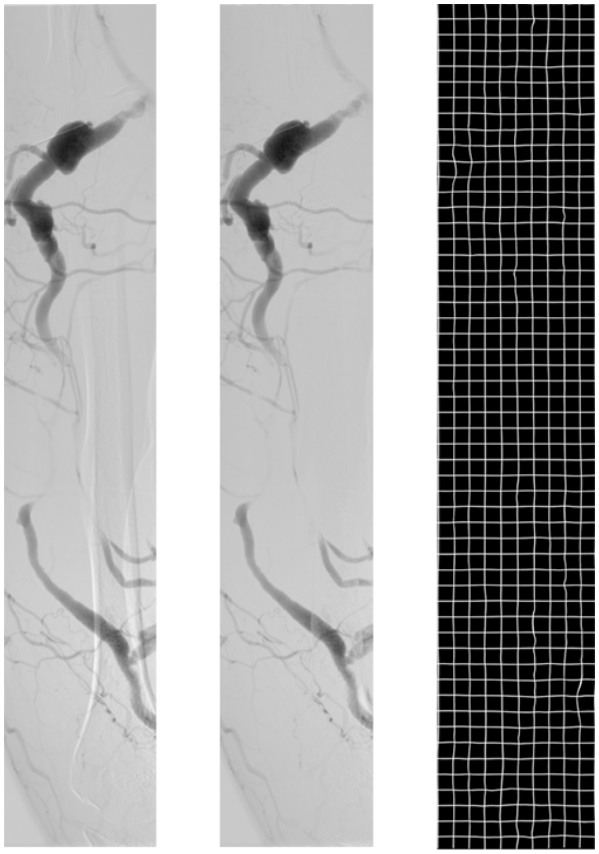


Figure 7: An image (a) before and (b) after registration with Elastix. (c) shows the corresponding deformation field of the mask image.

images of the spine and long bone images, allowing the vessels to be easily discerned.

Some drawbacks exist to the chosen method. Here accurate registration could not be achieved. Mainly these cases were abdominal images, like the image on the left. This is because simply too much movement occurs, and there is a large amount of contrast that is present in the image. In the image on the right, it appears that the registration has been unable to estimate shifts along the lower part of the image. This may be because window edges have fallen directly on the edge of the bone in that area. Although a sliding window has tried to limit this issue, unless the sliding window is extremely small, it is still possible to miss shifts in parts of an image where the edge of the bone and window for shift estimation are aligned. In the future, this method might be improved by avoiding placing window edges along the edge of the bone. Additionally, most of the time these artifacts appear near the image boundaries as there is not enough information to accurately estimate the shift in the area closer to the image boundary.

5.2. Virtual Subtraction

From Fig 10, it can be seen that the virtual subtraction has been quite successful in small patches. In many of the selected patches, very small vessels are found. This indicates that the U-net has had the desired high resolution output. Additionally, as shown in Fig 10.h, it can be seen that in cases where the groundtruth image contains no vessels, the virtual subtraction can recognize this. In many of the patches, though, some artifacts are found that were likely very bright patches in the live images, as in Fig 10.b, e, and g. In Fig 10.d, the vessel alongside the right part of the image is not found. This may be because the entire structure of the vessel was not able to be seen, and the network could not discern it from bone.

Virtual subtraction produces images of similar quality to the best DFT registration method. Very thin vessels are discerned, and motion artifacts are reduced. However, in some instances, the DFT registration gives a more accurate result. Because the intensity and shape of the contrast vary greatly between exams and frames, is thought that virtual subtraction struggles to discern the contrast from the background structures in these cases.

Virtual subtraction is an attractive alternative to full image DFT registration. Early on in this project, it was determined that pixel shifting was sufficient for removing artifacts in small areas, but could not be applied in larger regions. When computing many pixel shifts across an entire image and combining them, the registration improved. However, it was seen that some instances of DFT registration did not completely remove the artifacts. As images for training with virtual subtraction were generated in 224×224 sized windows, pixel shifting could be used, which created a nearly artifact-free training dataset.

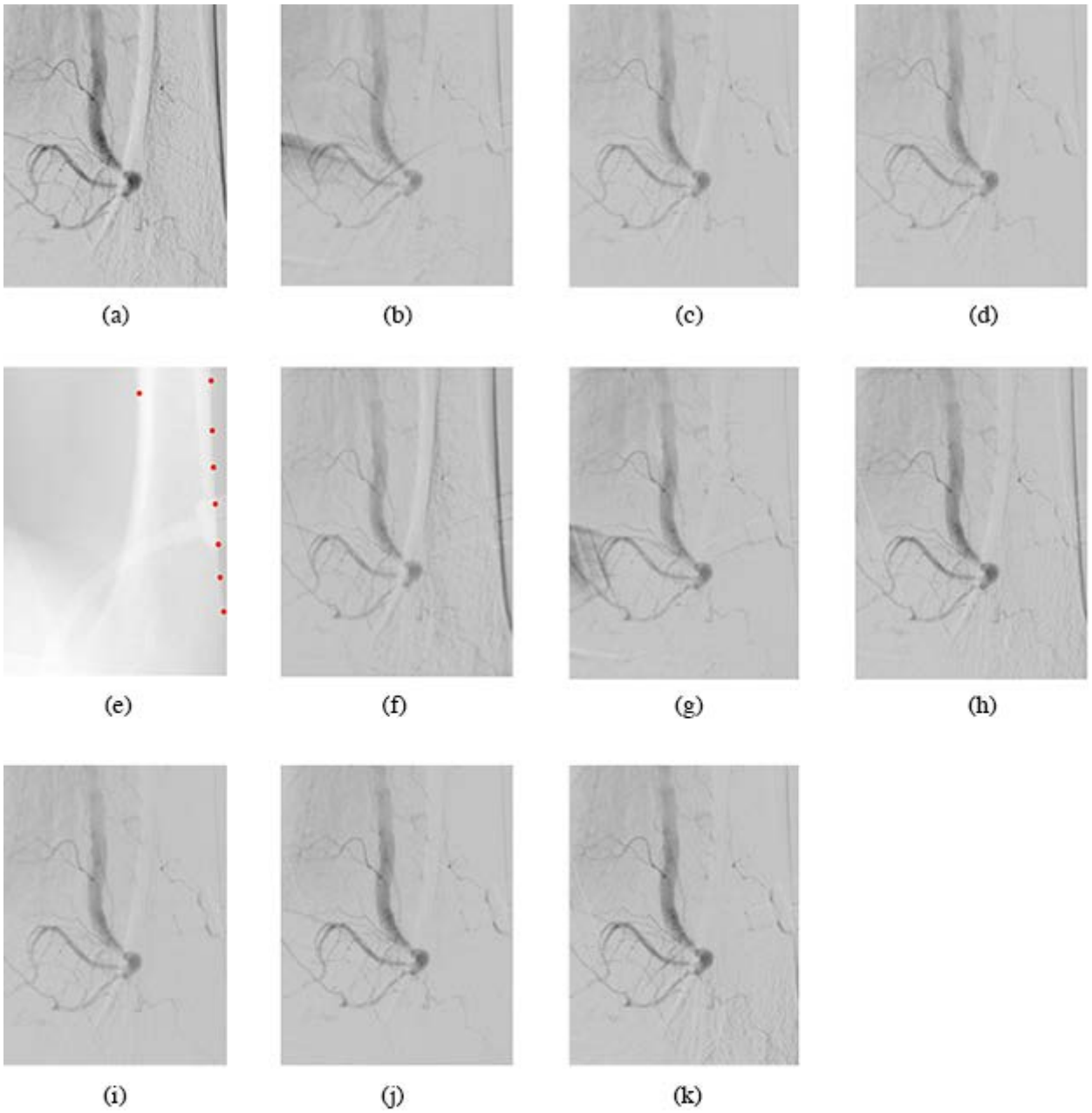


Figure 8: Comparing different methods for DFT registration. (a) is the unregistered subtracted image, (b)-(d) show Canny unsigned, Canny signed, and Gaussian derivative preprocessing, with window size 200×200. (e)-(f) compare the smart control point method: (e) shows smart control points superimposed on the mask image and (f) shows the result of DFT registration with those points. (g)-(k) compare various window size and spacings. (g) has window size 100 × 100 and (h) 300 × 300 with no overlap. (i-k) have window size 200 × 200 with window spacings 75, 100, and 150 respectively.

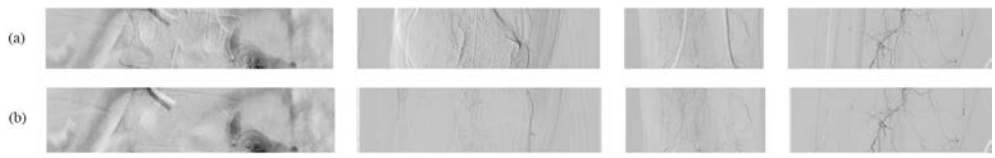


Figure 9: Successful DFT registration with Canny signed preprocessing, window size 200×200 and window spacing 150. (a) shows the unregistered subtracted images and (b) shows the registered.

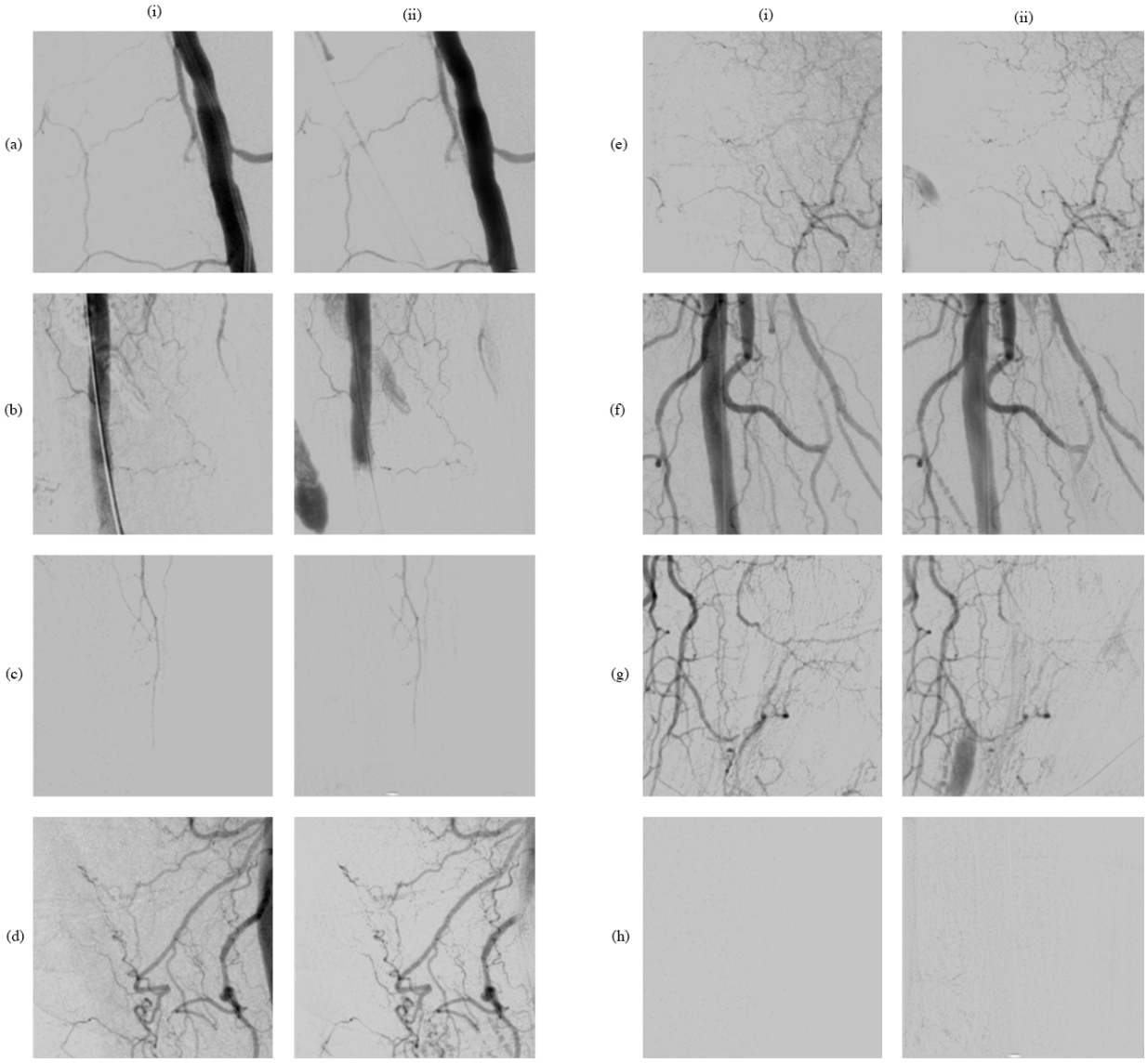


Figure 10: Virtual subtraction patches where column (i) corresponds to the image registered with DFT and (ii) to the image registered by virtual subtraction.

With many deep learning problems, issues often arise from the lack of data, and this problem is no exception. Perhaps during the initial data generation, a sliding window could be used to generate images with some overlap. Additionally, this dataset should be increased and carry a larger variety of images. For example, it is unknown how the model would perform on neuro- or cardio-angiography, which commonly require registration. With the proposed DFT technique, it can be assumed that the method could be applied successfully. Despite this, it has been shown that virtual subtraction has the potential in patient motion correction, especially in cases where a mask image is not available or the patient/detector motion is significant.

6. Conclusions

The findings of this project can be summarized in two main ideas. First, a pipeline for fast motion correction in full size DSA images based on the DFT has been proposed. This method can register 512×512 images accurately in less than one second. DFT registration has been shown to reduce patient motion in the subtraction image, making vessels appear more clearly in the areas of artifacts. In comparison to algorithms recently presented in the literature, faster performance has been achieved. DFT registration can be applied to a variety of images, such as exams of the head, legs, or knees. Secondly, the potential for virtual subtraction using a convolutional neural network has been explored. Virtual subtraction has the capability to widen the application of DSA in cases where conventional DSA is not available. Currently, virtual subtraction struggles to recognize certain structures satisfactorily, which may be due to the lack of training data of these structures. A future study could be conducted extending this method with a larger dataset, as well as testing the application of both DFT-DSA and virtual subtraction in neuro- and cardio-angiography cases.

7. Acknowledgments

I would like to thank my supervisor Eva Vandersmissen for providing the right balance of independence and guidance throughout my project and motivating me to consider that deep learning approaches do not trump, but have their place beside traditional image processing techniques. I would also like to thank all of my professors I have met throughout the MAIA master, who have given me the opportunity to learn so much over the two years. I can only express gratitude for all of my friends of the MAIA master, from every corner of the world, who have proven that snacks, laughter, and board games do brighten the most awful days. Lastly, I would like to thank my close friends and family for encouraging and supporting me throughout this whirlwind program.

References

Bentoutou, Y., Taleb, N., 2005. A 3-D space-time motion detection for an invariant image registration approach in digital subtraction angiography. Computer Vision and Image Understanding 97, 30–50. doi:10.1016/j.cviu.2004.07.002.

Crummy, A.B., Strother, C.M., Mistretta, C.A., 2018. The History of Digital Subtraction Angiography. Journal of Vascular and Interventional Radiology 29, 1138–1141. doi:10.1016/j.jvir.2018.03.030.

Eulig, E., Maier, J., Knaup, M., Koenig, T., Hörndler, K., Kachelrieß, M., 2019. Learned digital subtraction angiography (Deep DSA): Method and application to lower extremities. Proceedings of the 16th International Meeting on Fully 3D Image Reconstruction 1107223. doi:10.1117/12.2534740.

Fitzpatrick, J.M., 1988. The existence of geometrical density-image transformations corresponding to object motion. Computer Vision, Graphics, and Image Processing 44, 155–174.

Foroosh, H., Zerubia, J.B., Berthod, M., 2002. Extension of phase correlation to subpixel registration. IEEE Transactions on Image Processing 11, 188–199. doi:10.1109/83.988953.

Hiroshima, K., Funakami, R., Hiratsuka, K., Nishino, J., Odaka, T., Ogura, H., Fukushima, T., Nishimoto, Y., Tanaka, M., Ito, H., Yamamoto, K., 2001. Digital subtraction angiogram registration method with local distortion vectors to decrease motion artifact. Journal of Biomedical Informatics 34, 182–194. doi:10.1006/jbin.2001.1018.

Lee, S., Jeon, C.H., Sunwoo, L., Oh, D.Y., Lee, K.J., 2019. Phase-based nonrigid deformation for digital subtraction angiography. IEEE Access 7, 32256–32265. doi:10.1109/ACCESS.2019.2902562.

Liu, B., Jiang, Q., Liu, W., Wang, M., Zhang, S., Zhang, X., Zhang, B., Yue, Z., 2018. A vessel segmentation method for serialized cerebralvascular DSA images based on spatial feature point set of rotating coordinate system. Computer Methods and Programs in Biomedicine 161, 55–72. doi:10.1016/j.cmpb.2018.04.010.

Markelj, P., Tomaževič, D., Likar, B., Pernuš, F., 2012. A review of 3D/2D registration methods for image-guided interventions. Medical Image Analysis 16, 642–661. doi:10.1016/j.media.2010.03.005.

Meijering, E.H.W., Zuiderveld, K.J., Viergever, M.A., 1999. Image Registration for Digital Subtraction Angiography. International Journal of Computer Vision 31, 227–246.

Nejati, M., Pourghassem, H., 2014. Multiresolution image registration in digital x-ray angiography with intensity variation modeling. Journal of Medical Systems 38. doi:10.1007/s10916-014-0010-8.

Nejati, M., Sadri, S., Amirfattahi, R., 2013. Nonrigid Image Registration in Digital Subtraction Angiography Using Multilevel B-Spline. BioMed Research International 2013, 1–12. doi:10.1155/2013/236315.

Sailer, A.M.H., Grutters, J.P., Wildberger, J.E., Hofman, P.A., Wilmink, J.T., van Zwam, W.H., 2013. Cost-effectiveness of CTA, MRA and DSA in patients with non-traumatic subarachnoid haemorrhage. Insights into imaging 4, 499–507. doi:10.1007/s13244-013-0264-6.

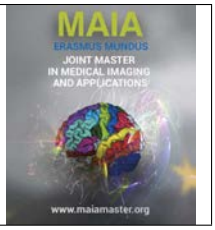
Unberath, M., Hajek, J., Geimer, T., Schebesch, F., Amrehn, M., Maier, A., 2017. Deep Learning-based Inpainting for Virtual DSA. IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC) , 1–2.

Wang, F., Guan, Q., Zang, P., Liu, X., Yang, T., Song, H., 2015. Is 3D Digital Subtract Angiography Really Perfect? Enlightenment from a Case with Both False Positive and False Negative Results. West Indian Medical Journal 26, 543–548. doi:10.7727/wimj.2015.141.

Zhang, X., Zhang, F., Li, R., 2010. DSA image registration based on 3D space-time detection. Procedia Engineering 7, 426–431. doi:10.1016/j.proeng.2010.11.070.

Medical Imaging and Applications

Master Thesis, June 2019



Deep learning methods for Image reconstruction (Super-Resolution)

Lavsen Dahal, Arnau Oliver, Xavier Llado

Computer Vision and Robotics Group, University of Girona, Catalonia, Spain

Abstract

High Resolution (HR) Magnetic Resonance Imaging (MRI) provides detailed anatomical information and is widely used in brain imaging diagnosis. However, HR imaging come at the cost of prolonged scans, high system cost and is subject to motion artifacts. Recently Single Image Super Resolution (SISR) techniques have gained a lot of attention due to their success with deep learning methods. Super Resolution MRI via deep learning has great potential clinical application as it will reduce scan time and improve patients comfort. It will also save costs as the same hardware that are already installed can be used to acquire images that are enhanced to have higher resolution by this method. In this paper, we propose two different Convolutional Neural Network (CNN) architectures for brain MRI super resolution, 3D-Light Super Resolution Network (3D-LSRN) and SR-UResNet. We validate 3D-LSRN for MRI images deblurring in the dataset of healthy subjects from a Catalonia hospital. We also show that our SR-UResNet architecture outperforms bicubic interpolation, and other CNN method for different downsampling methods, upto 4x less in all image planes in terms of visual quality and objective quality criteria such as Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index (SSIM). Further, we validate results by segmenting tissues and sub-cortical structures in the brain obtaining close overlap between HR and our predicted images.

Keywords: Super Resolution, Magnetic Resonance Imaging (MRI), Deep Learning, Convolutional Neural Networks (CNN)

1. Introduction

Magnetic Resonance Imaging is widely used for brain imaging since it is a non-invasive technique without any ionizing radiation, has superior soft tissues contrast and provides detailed anatomical information (Frahm et al., 1999). The image quality in MRI depends on different factors such as matrix size, field of view (FOV), slice thickness, signal to noise ratio (SNR) and magnetic field strength.

The spatial resolution in MRI is defined by the size of imaging voxels which could be different in the directions of imaging planes. The matrix size, FOV and slice thickness define the size of voxel and hence, the resolution. The matrix size is the number of frequency encoding steps and phase encoding steps in two different directions of the image plane. The phase and frequency encoding covers certain area that defines the FOV. The in-plane voxel size is computed by dividing FOV by the matrix size and increasing FOV in either direction increases the size of voxels and hence, decreases the resolution. The slice thickness defines the depth of the voxel. Moreover, the MRI scanners can have different voxel spacing for different planes. The voxel spacing can be isotropic with 1x1x1 mm or even anisotropic like 1x1x3 mm where through-plane voxel spacing is more than in-plane voxel spacing. The anisotropic setting of this type will require less samples and as a result, reduced scan times, but can miss details.

The MRI scanners also come with different magnetic field strength, ranging from 1.5T to 10T and beyond with higher magnetic field strength producing superior image quality (Anna Nowogrodzki, Nature, 2018). Fig. 1 shows the brain MRI of a human for 3T and 9.4T MRI scanner. The higher magnetic field scanners can capture minute details. However, they come at the cost of prolonged scans, causing patients discomfort, also, the scanners used to acquire high resolution images in

Figure 1: 3T (left) and 9.4T (right) brain MRI of human. Credit: Rolf Pohmann/Max-Planck-Institute for Biological Cybernetics

this way are more expensive. Thus, there is a trade-off to avoid prolonged scans which can induce the risk of patient motion and also patient discomfort, and so the scanners with magnetic field strength of 1.5T or 3T are generally used by hospitals today.

Recently super-resolution (SR) techniques have been proved to be effective method to enhance the resolution of the images. The goal of super-resolution methods is to use the lower resolution (LR) images to construct the corresponding high resolution (HR) images. The LR images are considered to be low pass filtered (blurred) and down sampled version of HR images. Thus, there is a loss of high frequency information due to down sampling process with low pass filtering. Henceforth, it is an ill-posed problem as multiple solutions exists when mapping from LR to HR space (Shi et al., 2016).
To generate HR images from LR images, certain image priors are exploited and based on image priors, SR algorithms can be categorized into four types, prediction models, edge based methods, image statistical methods, and patch based methods. The prediction models are interpolation based methods such as bilinear or bicubic which generate HR pixel intensities by weighted average of neighbourhood LR pixel values (Irani, 2009). In edge based methods, various edge features are used such as depth and width of an edge or parameters of gradient profile. These methods generate HR images with superior edges, however, are poor in modelling other high frequency structures like textures (Tai et al., 2010). Image statistical methods use heavy-tailed gradient distribution, sparsity property of large gradients, and total variation as regularization term to generate HR images (Sun and Shum, 2015). In patch based methods, the patches are cropped from HR and LR images, to learn the mapping functions. Different mapping functions are proposed in literature such as weighted average, kernel regression, support vector regression, gaussian process regression, sparse dictionary representation and recently convolutional neural networks (CNN). Moroever, different methods are proposed to blend the overlapping pixels such as markov random fields and conditional random fields (Yang et al., 2014).

In this work, super-resolution network using deep learning was successfully designed for natural 2D images and brain MRI images. The main contribution is on super resolution for brain MRI images. In summary, the work has the following achievements:

- We design a novel CNN architecture, we name it 3D-LSRN inspired from ResNet as first proposed by He et al. (2016) and DenseNets as first proposed by Huang et al. (2017). We make heavy use of short skip connections between the layer and the preceeding one like ResNet, however concatenate the feature maps for two successive layers. We validate the model successfully training for brain MRI super resolution.

- We also design another 3D-CNN model, we name it SR-UResNet, which shares U-Net like architecture as in the work of Ronneberger et al. (2015), however, is modified to have learnable parameters also at down sampling path and includes residual blocks at both encoder and decoder space of the network. The model is trained for super resolution of brain MRI images for different degradation type of HR images and upto 4x less resolution, we obtain superior value of PSNR and SSIM than traditional bicubic and spline interpolation methods and other CNN method. We also validate our results by the segmentation of brain tissues and sub-cortical structures and always get closer overlap for HR image and our prediction.

The remainder of the paper is organized as follows: Section 2 details current state of the art for both 2d natural images and 3d brain imaging MRI data. The material and methods are described in Section 3. Results are presented in Section 4, followed by discussions in Section 5. Finally, we conclude in Section 6.

2. State of the art

In this section, we describe the current state of art for super-resolution of images. For 2D natural images, only the recent algorithms are summarized which are all based on neural network architectures, as their performance has surpassed all previous traditional state of art results. However, for brain MRI images several approaches that are not based on deep learning are also included which we found relevant for this work. We detail the super-resolution of 2D natural images and 3D medical images in the following and subsequent paragraph respectively.

Super-Resolution Convolutional Neural Network (SRCNN) as in the work of Dong et al. (2016a) and faster SRCNN (FSRCNN) by Dong et al. (2016b)

showed promising results using simple CNN architecture for Single Image Super-Resolution (SISR) of 2D natural images. The first CNN architecture, SRCNN for super-resolution used only three layers. Since then, various CNN architectures have been proposed with increased network depth, with noteworthy improvements over SRCNN. For example, Kim et al. (2016) proposed VDSR network having 20 weight layers with residual learning approach for SISR problem. Lim et al. (2017) proposed EDSR network and showed superior results by removing the batch normalization module from conventional residual networks. Recently, Generative Adversial Networks (GAN) have gained lot of attention in super-resolution problem. Ledig et al. (2017) proposed GAN for SISR and showed almost indistinguishable results as the original HR image. Though PSNR, which is commonly used metric to assess the performance, was lower for the results, than state of the art, however, they demonstrated the results to be perceptually better, justifying that highest PSNR does not necessary reflect perceptually better SR result. Zhang et al. (2018) proposed RCAN to adaptively learn more useful channel wise features simultaneously with a network of over 400 layers for SISR. Though several works have been published for Single Image Super-Resolution using CNN approaches lately, there has been limited work on super-resolution for medical images domain.

Jog et al. (2014) used random forest (RF) approach to enhance the resolution for brain MRI images. Bahrami et al. (2016b) used Canonical Correlation Analysis (CCA) to enhance the quality of 3T image to look like 7T MRI using paired datasets scanned from same subjects and showed better segmentation results for brain tissues compared to tissue segmentation from 3T MRI. Bahrami et al. (2016a) subsequently used 4 layer CNN to generate 7T like images from 3T images utilizing priori of brain tissues with same dataset as aforementioned. Pham et al. (2017) used generative adversarial network for simultaneous high-resolution reconstruction and segmentation of brain MRI data. Chen et al. (2018b) used a 3D DCSRN for super-resolution of brain MR images, and subsequently mDCSRN by Chen et al. (2018a) with generative adversial network guided training producing significantly better quality SR images. Sánchez and Vilaplana (2018) used a 3D generative adversial network for super-resolution of brain MRI images with image gradients and mean square error as content term and least squares for adversial loss for the generator. Nie et al. (2018) used Fully Convolutional Network (FCN) with adversarial learning to generate 7T MRI images from 3T.

In most of the algorithms used for the SR task, the LR images are obtained by simulating the HR images. The HR images are gaussian blurred with different kernel width as requirement for downsampling. In case of MRI images, some methods also use Fast Fourier Transform (FFT) to convert the image data to Fourier domain, and remove the high frequency components. After removing the high frequency components, the data in Fourier space is mapped back to image space with inverse Fourier transform.

There has been limited work to generate 7T like MRI from 3T MRI, however, there has not been any work which generate 3T like MRI images from 1.5T MRI scanner to the best of our knowledge and the majority of the scanners used worldwide today are either 1.5T or 3T. Laurentius Huber (2018) initiative to map the locations of 7T MRI shows there are less than 100 such scanners in the world today . Moreover, when single scanner data sets are used, the simulation pattern to generate LR images from HR images are not uniform in the literature, and there is difficulty to compare the results.

3. Material and methods

In this work, different datasets were used for super resolution application. The 91-image dataset as in the work of Yang et al. (2010) was used for the training and Set5 by Bevilacqua et al. (2012) and Set14 by Zeyde et al. (2010) were used for testing and validation respectively for super-resolution of 2D Natural Images. The DIV2K dataset from NTIRE 2017 challenge by Agustsson and Timofte (2017) was also used for super-resolution of natural images. And for medical imaging data, two different datasets of brain MRI images were used. One of the dataset pairs came from two different Catalonian hospitals with different magnetic field strength of 1.5T and 3T of same subjects. We call it Dataset-I further in this work. Dataset-I consists of brain MRI volumes of 15 healthy subjects with no known history of brain realted disorders. And the other dataset, which we call Dataset-II, has 51 brain volumes with all subjects having Multiple Sclerosis(MS) disease. Only T1 weighted MRI scans were used for this work.

Our method for the 2D natural images is briefly discussed, followed by our methodology for 3D medical imaging data in this section. The first CNN method for super-resolution presented by Dong et al. (2016a) and subsequently the improved faster super-resolution network by Dong et al. (2016b) for 2D natural images was re-implemented and acquired similar results. As in the original implementation, the simulated lower resolution images were obtained by blurring with a Gaussian kernel, and sub-sample it by the upscaling factor. The first CNN method as in the work Dong et al. (2016a) upscaled the image by bicubic interpolation before feeding the image to CNN network, however, the subsequent faster model included upsampling in the network as Transpose Convolution Layer and we made the same choice. The original implementation was in Caffe framework and our implementation is in Keras framework with Tensorflow backend. EDSR network by Lim et al. (2017) which was NTIRE 2017 challenge winner was also trained and tested and obtained reported results

as by the author using the same source code that was made public (Sanghyun Son, 2017). The NTIRE challenge used DIV2K dataset by Agustsson and Timofte (2017), where the higher resolution images and corresponding lower resolution images were provided for 2, 3 and 4 downscaling factors.

For the medical dataset, we designed two different CNN networks, for two different datasets as described in section 3.2 and different types and levels of degradation performed on the original image to obtain LR simulated image. We also treated 1.5T to 3T MRI as super-resolution problem and trained CNN network. The organization of this section is as follows. We present the problem formulation in subsection, 3.1, data-preparation and downgrading methods in 3.2, CNN architecture in 3.3, training and experiments in 3.4 and evaluation criteria in 3.5.

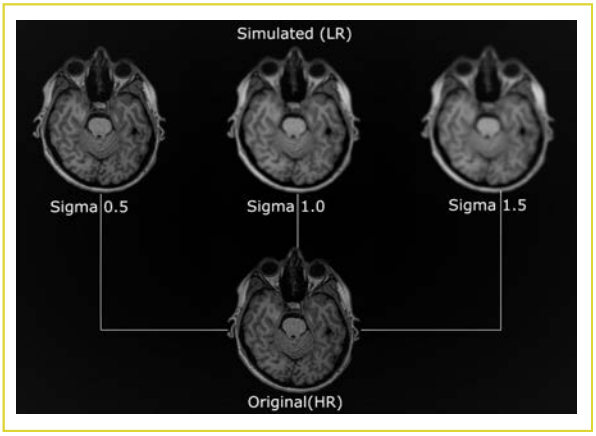


Figure 2: Original axial slice (bottom) of T1-weighted MRI of Dataset-I and simulated (top) obtained with Gaussian Blurring for different values of sigma

3.1. Problem Formulation

The LR and its corresponding HR image can be represented vectorially and symbolically denoted by y and x respectively. The relation between x and y is through some degradation model and can be presented as:

$$y = f(x) \tag{1}$$

where f is the degradation model. The aim of super resolution problem is to estimate x from y. It is an ill-posed problem as multiple solution exists to x for a given y. The reconstructed image $\hat{x}$ is obtained by minimizing the loss function:

$$\hat{x} = arg\,min_x\, f^{-1}(y) + \lambda R(x) \tag{2}$$

where, $f^{-1}(y)$ is the data fidelity term, R(x) is regularization term that provides certain image priors such as local spatial correlations, low rank or total variation and λ is regularization parameter. A common data fidelity term for super resolution application is mean square error or mean absolute error. In traditional super resolution methods, the regularization term R(x) is determined manually by extensive experimentation which is time consuming and can be inaccurate. However, with deep learning approach, feature extractions, non-linear mapping and image reconstruction, the three essential steps for image super-resolution is guided to learn from the network. The network during training learns to optimize the differences between the ground truth images and reconstructed images, and in the process extracts relevant features automatically. This gives neural network architecture state of the art performance in super-resolution tasks (Dong et al., 2016a).

3.2. Data Preparation and Downgrading Methods

In this work, we analyze brain MRI super-resolution in terms of magnetic field strength (1.5T LR to 3T HR), image quality and also spatial resolution. For super-resolution in terms of magnetic field strength, unique pair of dataset of same subjects scanned in 1.5T and 3T strength MRI scanner is used. The 1.5T MRI scans are co-registered to 3T magnetic scans using affine registration. The 1.5T MRI images are considered as LR and corresponding 3T scans are HR images.

For super-resolution in terms of image quality, the LR simulated images were obtained as Gaussian filtered version of HR with increasing value of Gaussian filter width - sigma. The sigma value used in increasing order of magnitude were 0.5, 1 and 1.5 for three different degradation levels. The Fig. 2 shows different levels of degradation on original HR slice for different sigma values.

For super resolution in terms of spatial resolution, we obtained the simulated LR images using two different techniques. The spline interpolation of order 3 and Fast Fourier Transform (FFT) is used to downsample to 2, 3 and 4 times lower resolution than the original ones. We downsampled to lower resolution in all three imaging planes or in slice selection, phase encoding and frequency encoding gradient direction in case of MRI image. For spline interpolation, the image is pre-filtered with third order spline filter before interpolation. The method for FFT transform is represented by Fig. 3. The image is transformed in FFT domain and the higher frequency components are removed from the original HR images based on downsampling factor and the low frequency k-space data is obtained. From the low frequency k-space data, the image is reconstructed simply by inverse Fourier transform which is LR simulated image. In Fig. 4, HR images and the corresponding lower resolution obtained using third order spline interpolation and truncating outer 3D k-space method are shown.

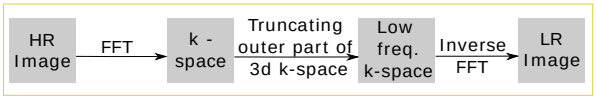


Figure 3: Block diagram for downsampling method using FFT

3.3. Convolutional Neural Network Architecture

We design two different CNN architecture for different downgrading methods. Our first model is light weight with only 152k trainable parameters and we call it 3D-LSRN (3D-Light Super Resolution Network). It has short skip connections like ResNet, however, we concatenate the feature maps instead of adding them. Thus, the following layer gets the feature maps of the last layer and, in addition, it also learns new feature maps. The second model is UNET like architecture with residual blocks and long skip connections. There is short skip connection in residual blocks and long skip connection between the input and the output and we call it SR-UResNet. We describe the light weight CNN architecture 3D-LSRN followed by SR-UResNet in this subsection.

3.3.1. Proposed 3D-LSRN (3D-Light Super Resolution Network)

We propose a new SR architecture for brain MRI image deblurring. The model is extremely light weight with only 152k trainable parameters making heavy use of (1x1x1) convolutions as first suggested in Lin et al. (2013) and we use it to reduce the dimension of feature space. We increase the feature maps dimension by concatenation and reduce it by (1x1x1) convolutions. Though (1x1x1) convolution is strictly linear, however, it is generally followed by non-linear activation function, exponential linear unit (ELU) in our case. This network architecture allowed us to go deeper and still have less trainable parameters. As we had only 15 brain volumes available in Dataset-I, and only 10 training volumes, bigger model would overfit. The network architecture 3D-LSRN is shown in Fig. 5 where, we have short skip connections after each convolution layer except for last two convolutional layers. The representation in latent space after each convolution is concatenated with the following layer. We have several advantages with this architecture design: light weight model and passing of feature maps at subsequent layer, less over-fitting as the number of training parameters are reduced.

The training is patch based, so the input to the network is the 3D patch of brain MRI images. The kernel size for convolution is (3x3x3) except for the reduction layer where the kernel size is (1x1x1). The exponential linear unit (ELU) is used as the activation unit except for the last convolution layer that maps the output. The last convolution layer has rectified linear unit (Relu) as activation function. The hyper-parameters and optimizer selection, training procedure and implementation framework are discussed in Section 3.4.

3.3.2. Proposed SR-UResNet Architecture

We propose another architecture for super resolution in terms of spatial resolution. We designed U-Net like architecture but also included residual blocks. There are no max pooling layers, instead convolution of stride 2 to perform downsampling in the network path. This method extracts more relevant features as the model has learnable weights for downsampling unlike non-learning layer like max pooling. The residual block is usual as in He et al. (2016) with batch normalization layers. The kernel size is (3x3x3) for all layers except for the last layer and the deconvolution layers. The kernel size for deconvolution layer is (2x2x2) as we noticed checkerboard artifacts in our reconstructed images when there was uneven overlap as suggested in (Odena et al., 2016). Deconvolution layer has uneven overlap when the kernel size is not divisible by the stride. As the stride for deconvolution layer is 2, we used kernel size of 2 in all deconvolution layers that is in the decoder path of the network, to not have uneven overlap, and consequently, our predicted image volumes are free of checkerboard artifacts.

The network architecture is shown in Fig. 6. The training is patch based, for this architecture as well, and the input to the network is (32x32x32) brain MRI patches. The network consists of convolution blocks, residual blocks and convolution of stride 2 in the encoder path of the network. Each convolution block has 2 convolution layers. One of the convolution layer has linear activation and the other one has non-linear activation as ELU. The residual block is similar to convolution block but in addition have short skip connections like ResNet. The batch normalization is performed in both convolution and residual block.

Let us represent the number of filters in convolution layers by n_f. From the fig. 6, it can be observed, we have four levels in the network as we used three convolution layers of stride 2 after every residual block in the encoder path. We began with (32x32x32x1) and our latent representation is (4x4x4xn_f) as we reached the bottom of the encoder path.

In the decoder path of the network, there is a deconvolution layer at each level followed by residual blocks. There are also skip connections between encoder and decoder which are at same levels except on the last level and a long skip connection between the input and the last convolution layer. As with the super-resolution application, having a long skip connection proved to be effective as the network has direct path between input and the last convolution layer. The addition of the input with the output of last convolution layer is the output or the prediction.

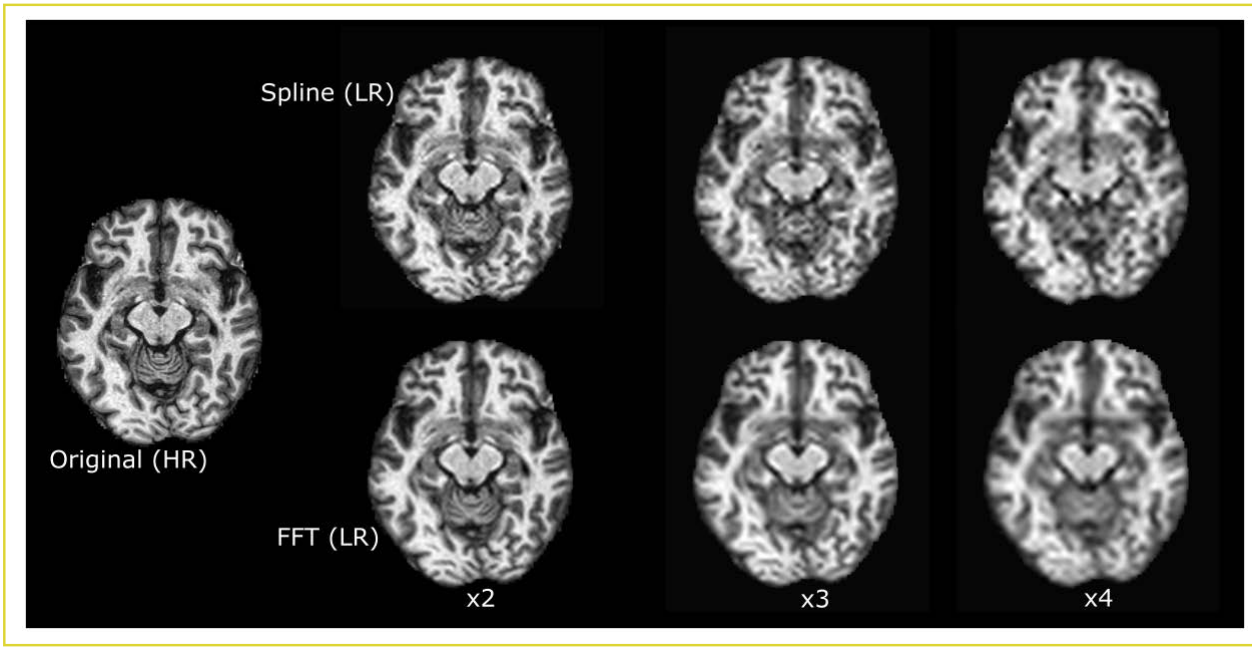


Figure 4: Downsampling with cubic spline interpolation and FFT. x2, x3 and x4 represents 2, 3 & 4 times lower resolution images than the original HR image respectively

Table 1: List of experiments with Dataset-I and model 3D-LSRN

Exp No.	Input	Output
1.1	Gauss Filtered - Sigma 0.5	3T MRI
1.2	Gauss Filtered - Sigma 1.0	3T MRI
1.3	Gauss Filtered - Sigma 1.5	3T MRI
1.4	1.5 T MRI Co-registered	3T MRI

3.4. Training and Experiments

We describe the training strategy and experiments for 3D-LSRN followed by SR-UResNet model. The 15 brain volumes from Dataset-I were randomly distributed into 10 training, 3 validation and 2 testing volumes. We performed four experiments with this model. The three experiments were with 3T MRI images and different simulated images were obtained. And the last experiment with this model was with dataset pair of 3T and 1.5T MRI where 1.5T MRI is LR and 3T is HR. As both 3T and 1.5T MRI images were from same subjects, they were co-registered affine to bring it in the same space. The table 1 lists different experiments with this model.

We used Gauss blurring for sigma values - 0.5, 1 and 1.5 and obtained three different sets of LR simulated images. The degraded images were the input and original 3T MRI images were the ground truth. In experiment 1.4, we treated 3T brain MRI image as HR and co-registered 1.5T MRI image as input. Isotropic patches of 32 voxels were extracted both from the input and the ground truth and fed to the network. We did not perform any kind of pre-processing other than re-scaling the image between -1 to 1 for this 4 experiments with Dataset-I. The brain images were with skull without any noise correction.

The model was trained with adam optimizer and mean absolute error loss function was used. A very low learning rate of 10e-5 was used. The model was trained for around 200 epochs. The non-activation function used was ELU. The learning rate was reduced by half when the loss plateaued. ELU was found to have given better results than ReLu activation function in terms of convergence speed as expected. This behaviour of different activation functions are described with our training strategy for SR-UResNet. For validation, additional metrics such as PSNR and SSIM were observed. The best performing model in validation set was selected as trade-off among PSNR, SSIM and mean square error as metrics. The model was implemented in Keras framework with tensorflow backend.

We use Dataset-II for SR-UResNet model for our experiment with super-resolution in terms of spatial resolution for brain MRI images. As this dataset, had 3 times more brain volumes than Dataset-I, and also consisted of subjects with Multiple Sclerosis, we chose this dataset for different experiments as it reflected potential real case application. The dataset was randomly distributed in the ratio 75% in training set and 12.5% each in test and validation set. The original image or HR is degraded either using FFT or cubic spline interpolation for 2, 3 and 4 times lower resolution in all 3 image planes for obtaining the simulated LR image. As we observed checkerboard patterns when increasing the resolution through CNN model, we increase the resolution by spline interpolation of order 3 for both FFT

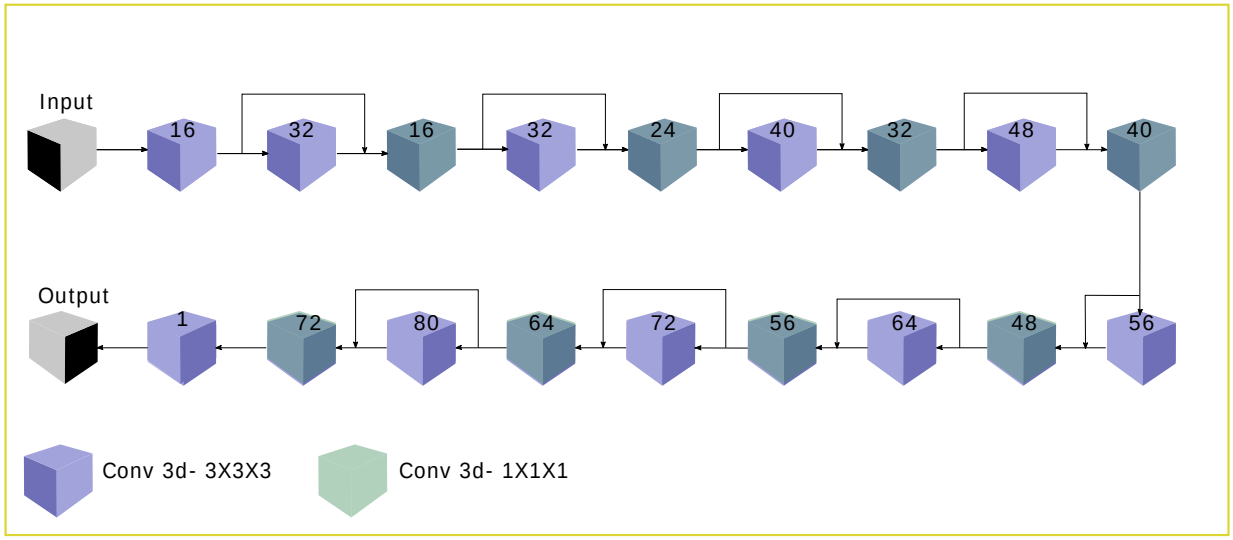


Figure 5: Architecture for 3D-LSRN. The purple and green colour denotes the 3D Convolution with kernel size (3x3x3) and (1x1x1) respectively. The number on the top of the cube indicates the number of filters in that layer.

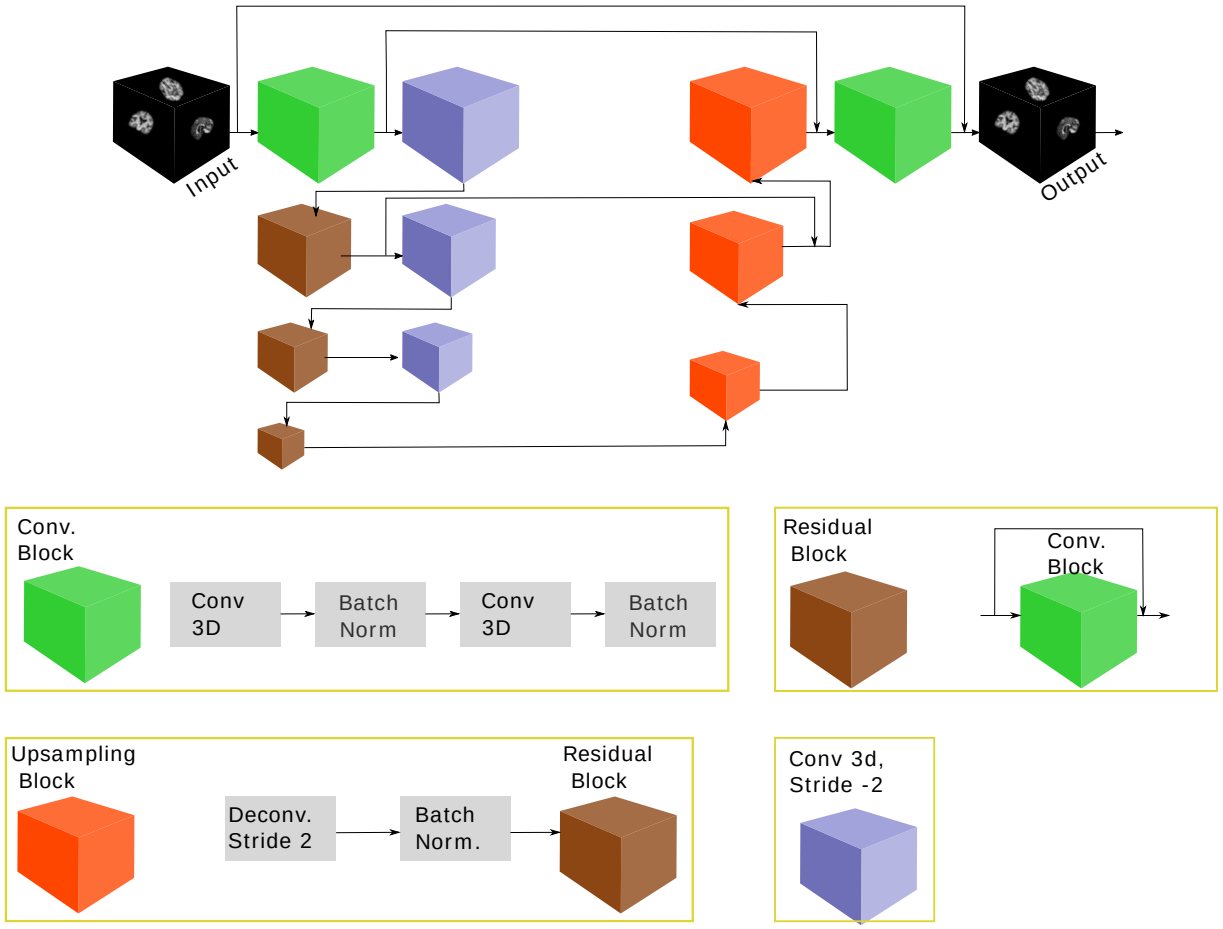


Figure 6: Architecture for SR-UResNet. The input to the network is 32x32x32 size patch. The green, brown and orange colour cube denote the convolution block, residual block and upsampling block respectively. The residual block is simply convolution block with skip connections. For upsampling, we use Deconvolution of stride 2, and upsampling block has Batch Normalization layer and residual block as well. The purple colour cube denotes the Convolution 3D which has stride 2 that is used to downsample the network path.

Table 2: List of experiments with Dataset-II and model SR-UResNet

Exp No.	Degradation Method	Degradation Level	Up-scaling Method	Up-scaling Level
2.1	cubic spline	x2	cubic spline	x2
2.2	cubic spline	x3	cubic spline	x3
2.3	cubic spline	x4	cubic spline	x4
2.4	FFT	x2	cubic spline	x2
2.5	FFT	x3	cubic spline	x3
2.6	FFT	x4	cubic spline	x4

and spline degraded image so the input has same spatial resolution as ground truth or output. The table 2 lists the experiments for different degradation levels and types. We used the simulated images generated by cubic spline interpolation of two times lower resolution as input and HR images as ground truth and tuned the hyper-parameters for the model. The hyper-parameters giving the best results for this dataset pair in terms of PSNR, SSIM metric and mean square error are used for all the other experiments using model SR-UResNet. Skull stripping was done as we also wanted to evalulate the segmentation results for both tissues and subcortical structures of brain as one of our evaluation criteria for this model. We expected the Dice similarity Coefficient for segmentation results of HR with our prediction image to be better than HR with LR images. We show in the results that we always get better Dice for our prediction images with HR than LR with HR, justifying the model has learnt to enhance the resolution without artifacts and also preserving minute structural details.

We used ELU as activation function as we had dying ReLU issue when using ReLU activation function. A dying ReLU problem can be caused by a large gradient flowing through a ReLU neuron and always outputs same value - 0. As when ReLu ends at this state, it is unlikely to recover as the gradients of 0 is 0, and the weights are not optimized further. We use ELU as activation function Clevert et al. (2015) to fix this issue. The ELU unit can be represented as:

$$ELU(x) = max(0, x) + min(0, \alpha * (exp(x) - 1)) \quad (3)$$

where the Elu hyperparameter α controls the value to which an ELU saturates for negative net inputs. The Fig. 7 shows the Elu activation function for α =1, where it can be observed that Elu unit allows small gradients to flow in negative direction as well and saturates based on given α. For this reason, Elu diminish the vanishing gradient effect seen with ReLu unit. For these experiments, the value of α used is 1.

For the loss function, we use Huber loss as suggested in (Girshick, 2015) which is given by:

$$loss(x, y) = 1/n \sum_i z_i \quad (4)$$

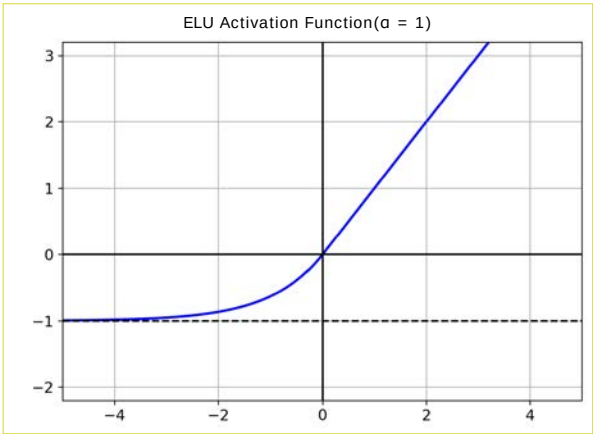


Figure 7: Graph for ELU activation function. Unlike to ReLu, ELU can produce negative outputs.

where z_i is given by:

$$z_i = \begin{cases} 0.5(x_i - y_i)^2, & if \mid x_i - y_i \mid < 1 \\ \mid x_i - y_i \mid -0.5, & otherwise \end{cases} \quad (5)$$

where x_i and y_i are the vectorial representations of the input and prediction respectively. This loss is less sensitive to outliers than MSE loss (Girshick, 2015) and we had better performance with this loss function. This model was implemented in Pytorch framework. Both the models were trained in NVIDIA GTX-1080 GPU with 12GB memory and 128GB RAM.

3.5. Evaluation Criteria

For evaluation of our results, we used subjective and objective evaluation. The subjective evaluation is based on visual perception of human eyes to evaluate image quality. However, this evaluation can be different for individuals. For objective evaluation, we computed PSNR, SSIM and also segmented the brain tissues and sub-cortical structures to compute Dice Similarity Coefficient (DSC).

We used PSNR, SSIM, mean square error as evaluation metrics during training and validation. PSNR and SSIM are commonly used metrics to evaluate the super resolution performance. PSNR is given by:

$$PSNR = 10log_{10}\left(\frac{R^2}{MSE}\right) \quad (6)$$

where R, is the maximum fluctuation of input data type. For example, if the input data type has an 8-bit unsigned integer data type, R is 255.

SSIM index is based on the computation of three terms, luminance term, contrast term and structural term. The overall index is multiplicative combination of the three terms. Let l(x,y), c(x,y) and s(x,y) denote the luminance term, contrast term and structural term respectively. Then, SSIM as first suggested in Wang et al. (2004) is given by:

$$SSIM(x, y) = [l(x, y)]^{\alpha}.[c(x, y)]^{\beta}.[s(x, y)]^{\gamma} \quad (7)$$

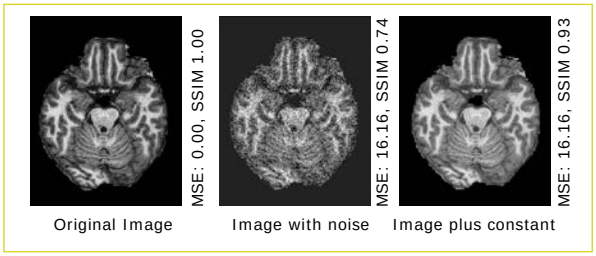


Figure 8: Comparison of SSIM and MSE as a metric. The image with noise and image plus constant has same MSE, however different SSIM. As SSIM also takes into account texture information, it performs better than MSE.

where,

$$l(x,y) = \frac{2\mu_x\mu_y + C1}{\mu_x^2 + \mu_y^2 + C1} \tag{8}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + C2}{\sigma_x^2 + \sigma_y^2 + C2} \tag{9}$$

$$s(x,y) = \frac{\sigma_{xy} + C3}{\sigma_x\sigma_y + C3} \tag{10}$$

where μ_x, μ_y, σ_x, σ_y, and σ_{xy} are the local means, standard deviations and cross-variances for images x and y and C1, C2 and C3 are constants for luminance term, contrast term and structural term respectively. SSIM is better indicative than just mean square error as metric as it also accounts for texture as in the work of Wang et al. (2004) We explain it with the fig. 8 as an example. The input image is modified by adding random noise in first case and by adding a constant in the second case. In both the cases, mean square error is the same, however, SSIM values are different.

We also evaluated our performance using the segmentation of brain tissues and sub-cortical structures. Both tissues and sub-cortical structures are segmented using two independent software packages, volbrain (Manjón and Coupé, 2016) and FSL. FAST (Zhang et al., 2001) package from FSL was used for tissue segmentation and FIRST (Patenaude et al., 2011) was used for sub-cortical structures segmentation. Dice Similarity was used as metric to evaluate the segmentation performance. The results for segmentation are shown in section 4 both quantitatively and qualitatively.

4. Results

In this section, results for super-resolution using both of the proposed architecture are presented in the following order:

- The results for experiments 1.1 - 1.3 where LR images were simulated as Gaussian blurred image of HR for different magnitude of kernel width trained with 3D-LSRN model and Dataset-I. PSNR and SSIM are presented as quantitative metrics and actual predictions along with the original and blurred images are shown for qualitative evaluation.

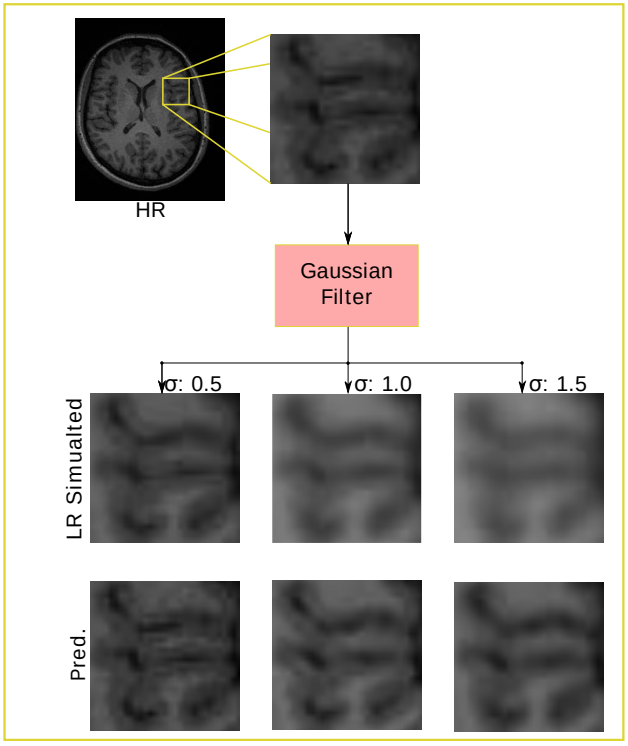


Figure 9: Qualitative Results for Dataset-I with trained with 3D-LSRN model. The HR image axial slice and a particular zoomed in region of HR and corresponding region for LR simulated images obtained by applying a Gaussian Filter for different values of sigma and the predictions are shown.

- The results for experiment 1.4 where HR images were 3T MRI and LR images were 1.5T of same subjects co-registered in affine space trained with 3D-LSRN model and Dataset-I. The evaluation metrics presented are same as above.

- The results for experiments 2.1 - 2.6 where LR images were simulated by downgrading up to 4x lower resolution in all image planes by two different downsampling operators trained with SR-UResNet model and Dataset-II. In addition to all evaluation criteria used for 3D-LSRN model, the segmentation for brain tissues and sub-cortical structures are presented qualitatively and DSC is reported.

In Fig. 9, we visually show that our model 3D-LSRN consistently yields higher quality image for different levels of image blurring and in Table 3 we present quantitative results. From Table 3, it can be noted that PSNR for our prediction and HR is atleast 10dB more than PSNR for LR and HR for all degradation levels. In Fig. 10, we present our results when 3D-LSRN model was trained with 3T MRI as HR and 1.5T as LR. The average PSNR and SSIM for test set between prediction and ground truth for this configuration were computed as 21.80 and 0.56 respectively.

In Fig. 11, results for Dataset-II trained with SR-

Table 3: Quantitative Results for Dataset-I with 3D-LSRN Model

Sigma	Image	PSNR	SSIM
0.5	Input	29.41	0.972
	Pred.	45.23	0.997
1.0	Input	20.83	0.843
	Pred.	32.16	0.943
1.5	Input	19.11	0.769
	Pred.	29.35	0.912

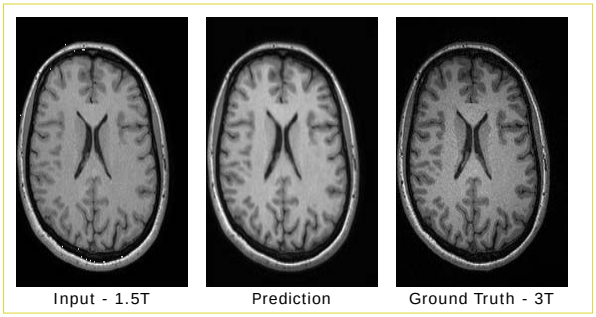


Figure 10: Qualitative Results for 3D-LSRN trained with 3T images as ground-truth and 1.5T as input. From left to right (input that is is the image from 1.5T scanner co-registered to the image from 3T scanner of same subject, prediction from 3D-LSRN, ground truth which is from 3T scanner.)

UResNet model is presented qualitatively where LR images were obtained by down sampling to 2,3 and 4 times in all image planes by two different downsampling operators - cubic spline interpolation and truncating outer 3d k-space. The axial slice of HR image and zoomed in small region of that slice are shown qualitatively for all degradation levels along with prediction. The complete downsampled axial slice and corresponding prediction along with original HR are shown in Fig. 13.

In Fig. 12, quantitative results in form of boxplots for evaluation metrics PSNR and SSIM are presented. Both PSNR and SSIM have a low variance. Even for 4x downsampling in all imaging planes, 0.93 for SSIM and 31.64 for PSNR when downsampled using spline interpolation and 0.94 for SSIM and 34.63 for PSNR when down sampled truncating k-space are obtained. The objective metrics obtained by our method is compared with bicubic interpolation, cubic spline interpolation and FSRCNN as in the work of Dong et al. (2016b) in Fig. 12. As FSRCNN model was implemented for 2D natural images, it was re-implemented in 3D with same hyper-parameters proposed by the authors. Our method obtained superior SSIM and PSNR values for all downgrading levels and approaches.

In Fig. 14 quantitative results in form of box plots for DSC between segmentation of HR and prediction of our method - SR-UResNet and also between segmentation of HR and cubic spline interpolated result are shown. The DSC for prediction of our method and HR is always superior than cubic spline interpolated results for all tissue types and sub-cortical structures for all downsampling levels and operators.

In Fig. 15, qualitative results of segmentation of HR images and predictions from our method are presented. The segmentation is presented as overlay on the coronal slice of the HR image volume, cubic spline interpolated result, and the prediction from our method (SR-UResNet).

It can be observed in Fig. 15 that segmentation for our prediction images are much smoother than corresponding LR simulated images and for the simulated images obtained by cubic spline interpolation downsampling operator for 4 times lower resolution, Left thalamus is not segmented at all, and almost negligible right thalamus is segmented. However, for the prediction image from model SR-UResNet, even for 4x less lower resolution, both left and right thalamus have been segmented with superior overlap. Even for downsampling by truncating outer 3D k-space, there is greater overlap between our prediction and HR than cubic spline interpolated image and HR. There is over segmentation for LR simulated images obtained with downsampling operator - truncating outer 3D k-space for 3 and 4 times lower resolution possibly due to loss of sharp contrast edges by removal of high frequency components. The segmentation is obtained from volbrain (Manjón and Coupé, 2016).

In Fig. 16, the tissue segmentation for brain tissues is shown for HR axial slice and simulated LR and prediction. The image that is downsampled to the lowest resolution that is 4 times in all image plane is shown. In Fig. 16, it can be observed segmentation of CSF is superior in our prediction image than LR image obtained by both downsampling operator. Also, the segmentation for LR image obtained by cubic spline interpolation for CSF is coarse, however the CSF segmentation for our prediction image is smooth and has superior overlap with ground truth.

5. Discussion

In this study, two different CNN architectures for super resolution of brain MRI images have been proposed and implemented. The LR images were obtained by different downgrading methods like Gaussian blurring, cubic spline interpolation, and truncating outer 3D k-space. The first proposed CNN architecture 3D-LSRN was validated for MRI image deblurring with LR images obtained by degrading HR images by Gaussian filtering of different kernel width. The prediction images obtained for LR images obtained by Gaussian blurring with sigma 0.5 were visually indistinguishable compared to HR. The increment for PSNR was more than 16dB between the degraded image and HR and network prediction and HR and SSIM was computed as 0.997 in the test set for this downgrading configuration. The model successfully learnt to deblur the images. Even for

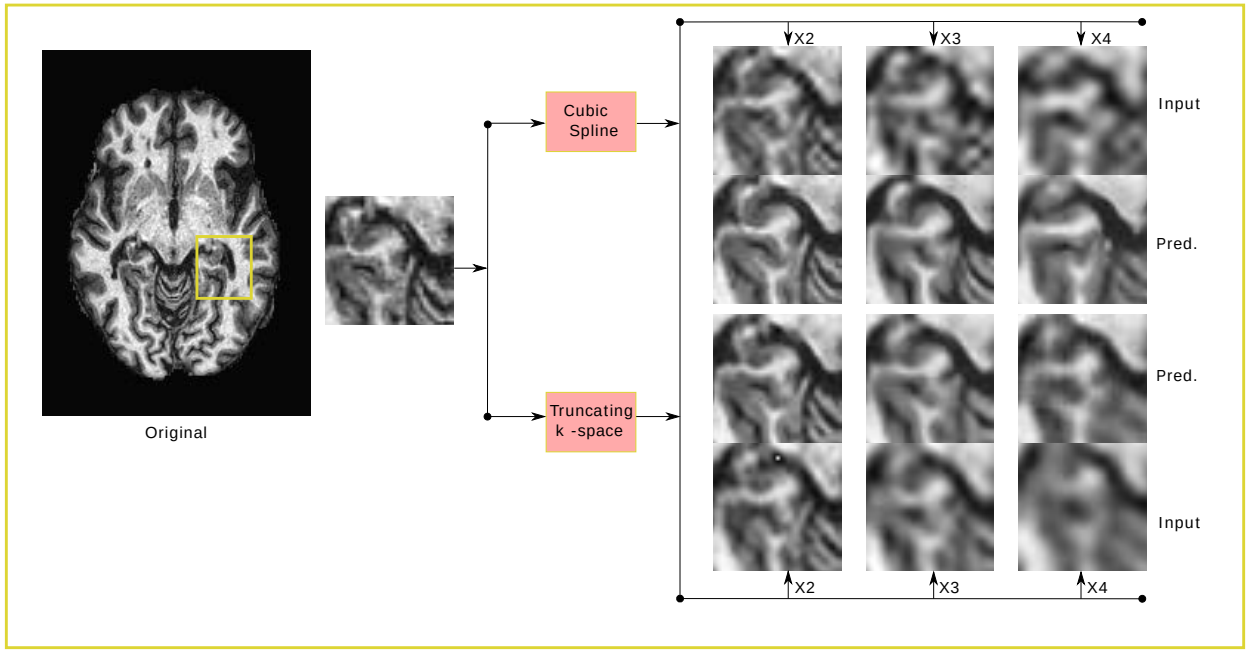


Figure 11: Qualitative Results for SR-UResNet model trained in Dataset-II. The original images are degraded by two methods: cubic spline and truncating k-space, for 2, 3 and 4 times lower resolution. The simulated LR and prediction for a region of interest marked by a yellow color square in original image for all down-sampling methods and all levels are shown. x2, x3 and x4 represents 2, 3 & 4 times lower resolution images than the original HR image respectively.

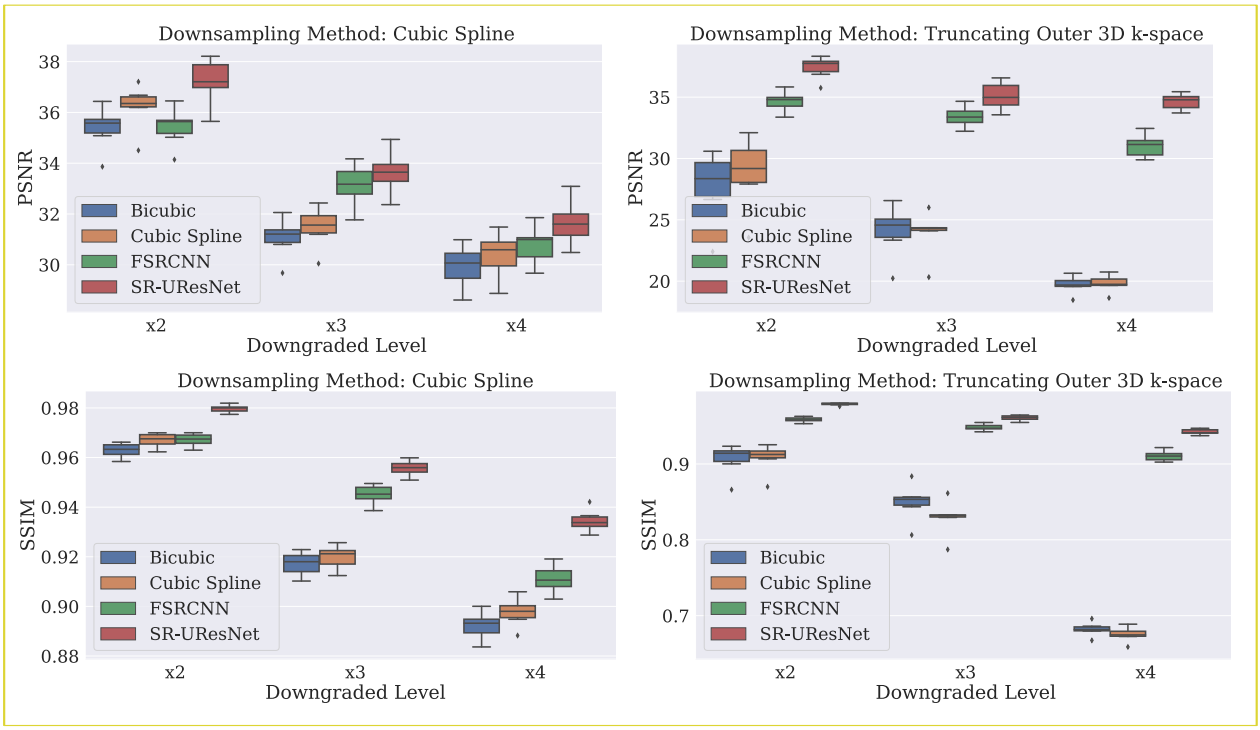


Figure 12: Comparison of PSNR and SSIM with other methods for both downsampling strategies - cubic spline and truncating outer 3D k-space . The boxplots for PSNR and SSIM are shown in first and second row respectively. x2, x3 and x4 represents 2, 3 & 4 times lower resolution images than the original HR image respectively.

degraded image with higher magnitude of kernel width, PSNR and SSIM were reported much higher for prediction of the network and HR than corresponding LR images and HR as shown in table 3. The 3D-LSRN network was trained for other experiment where HR images were 3T and LR images were 1.5T that were co-registered as they were of same subjects in both scanners. However, the results were underwhelming possi-

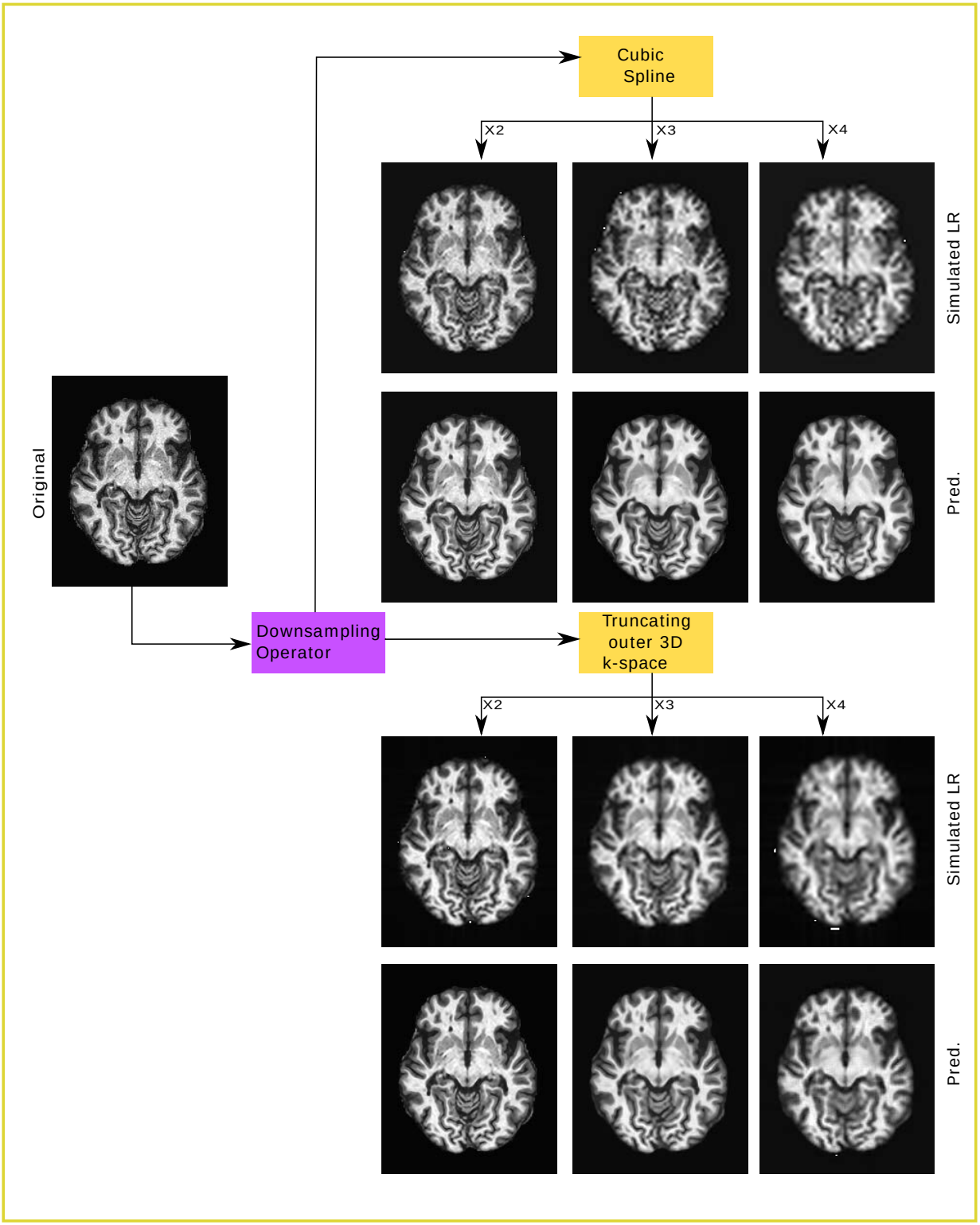


Figure 13: Qualitative Results for SR-UResNet Model. The original axial slice which is downsampled by two different methods - cubic spline and truncating outer 3d k-space and the corresponding predictions are shown. x2, x3 and x4 represents 2, 3 & 4 times lower resolution images than the original HR image respectively.

bly due to small dataset size of only 10 training volumes.

In another experiment, 1.5T MRI images were tested with the model that was trained with 3T MRI as groundtruth and simulated LR images (obtained by downsampling operator - Gaussian Blurring with sigma 0.5 and 1.0) as input. From Fig. 17, it can be observed that when tested with model which was trained with

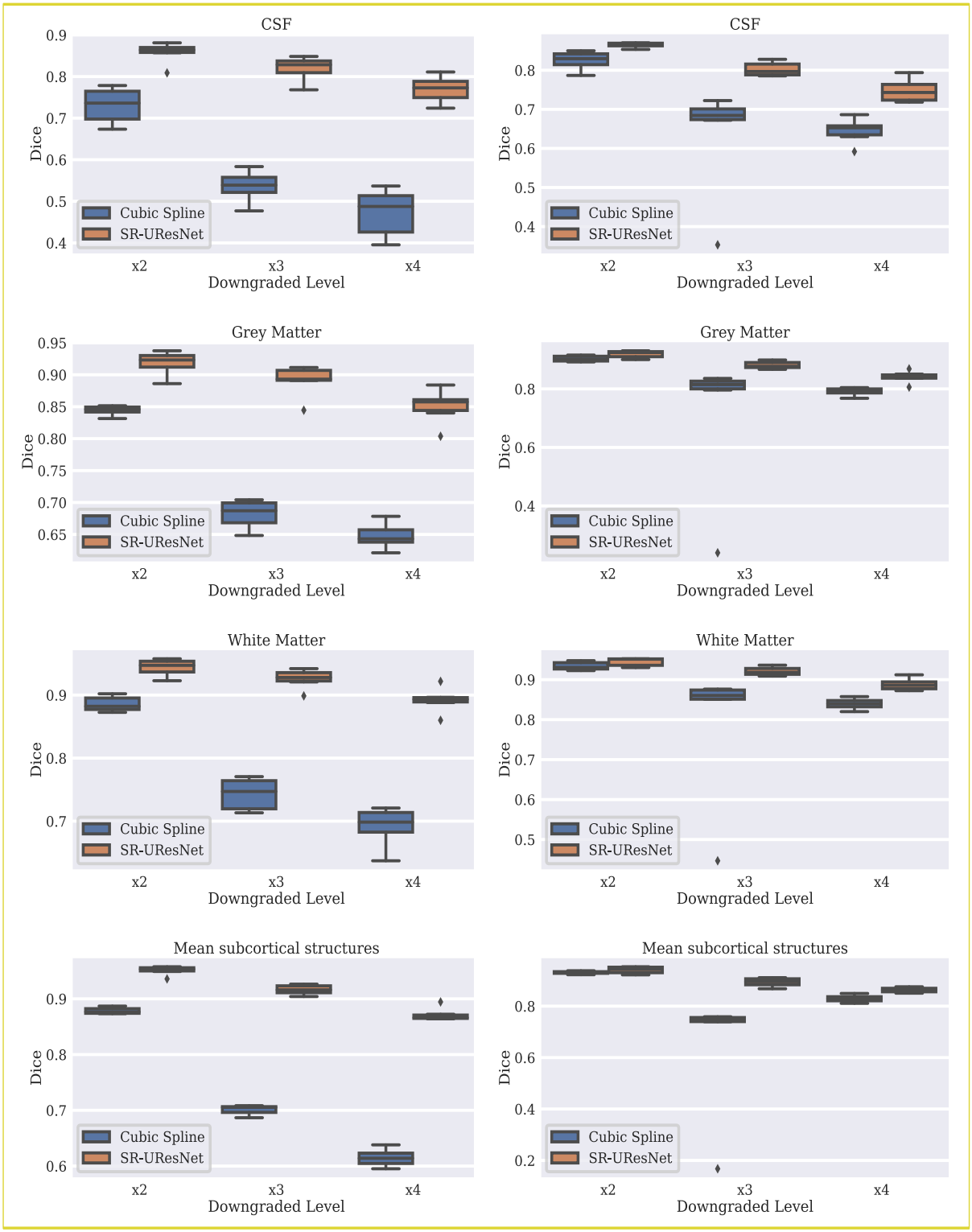


Figure 14: Boxplots for Dice Similarity Coefficient (DSC) for brain tissues - CSF, Grey Matter (GM) and White Matter (WM) and subcortical structures. The first and second column in every row represents the DSC boxplot for the tissue or sub-cortical structures as observed in the title of the boxplot for LR images obtained with downsampling operator truncating outer 3D k-space and cubic spline respectively. x2, x3 and x4 represents 2, 3 & 4 times lower resolution images than the original HR image respectively.

simulated image obtained with sigma 0.5, the prediction looked much similar to the ground truth than when tested with simulated image obtained with sigma 1.0 possibly due to simulated image obtained with sigma 0.5 was more similar to 1.5T MRI than simulated image obtained with sigma 1.0. This was interesting insight as

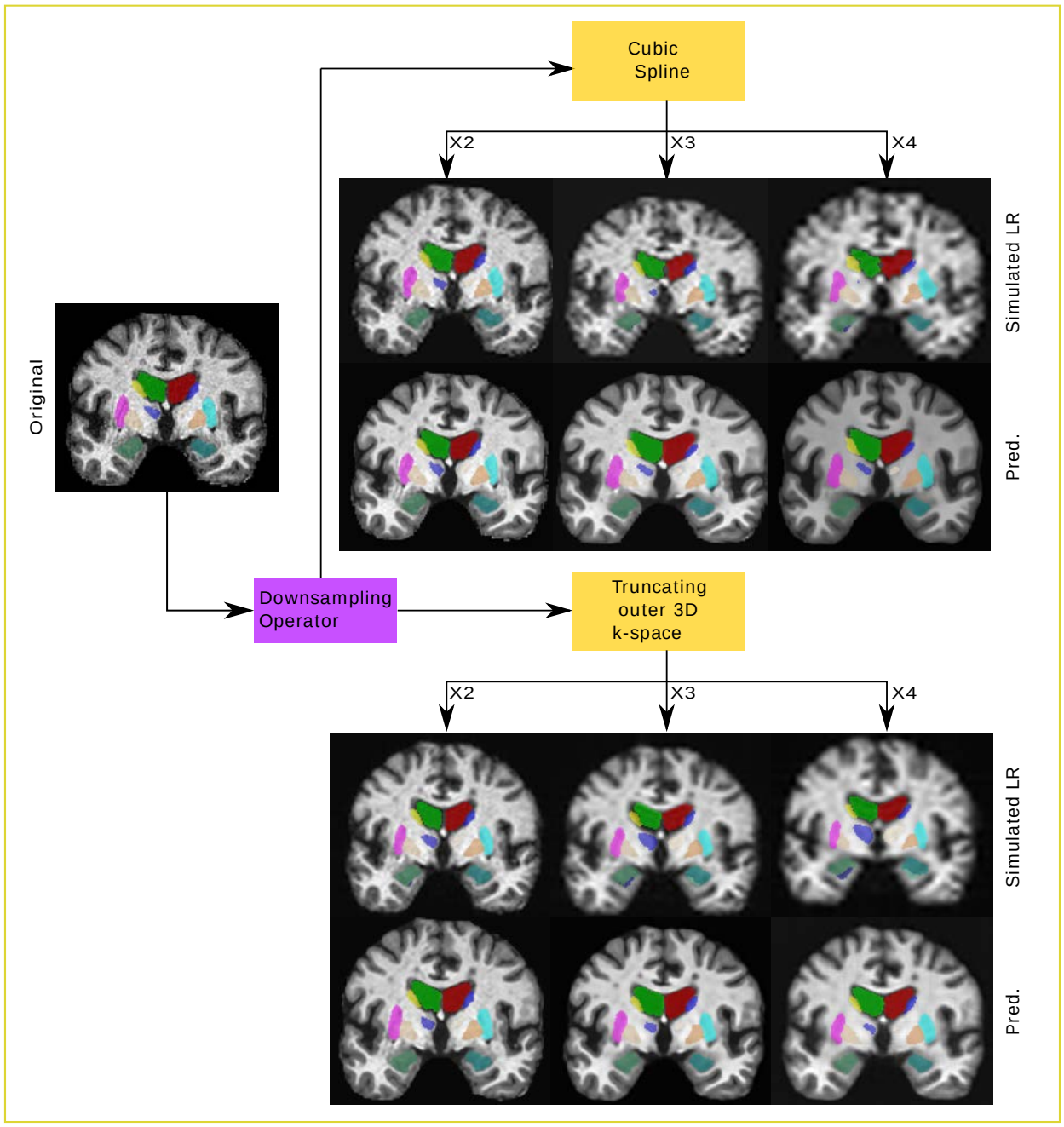


Figure 15: Segmentation of sub-cortical structures for HR, LR simulated images and prediction from SR-UResNet model. The segmentation is shown as overlay in the corresponding coronal slice. x2, x3 and x4 represents 2, 3 & 4 times lower resolution images than the original HR image respectively

the model learnt to transfer style.

In other set of experiments with SR-UResNet model, LR images were obtained by downsampling with two different downsampling operators, cubic spline interpolation and truncating outer 3D k-space. Even for 4x lower resolution simulated images, the predictions from SR-UResNet were very similar to ground truth visually. PSNR and SSIM were always superior than traditional methods like bicubic and other CNN method as shown in boxplots in Fig. 12. The prediction of our network were further validated by segmenting the tissues and sub-cortical structures. DSC between ground truth and our prediction for all downsampling strategies and levels were always higher than cubic spline interpolated result and ground truth. The prediction images obtained were without any artifacts and minute structural details were preserved which is validated visually and also supported by high DSC computed for segmentation of different sub-cortical structures. The mean DSC for sub-cortical structures segmentation increased by more than

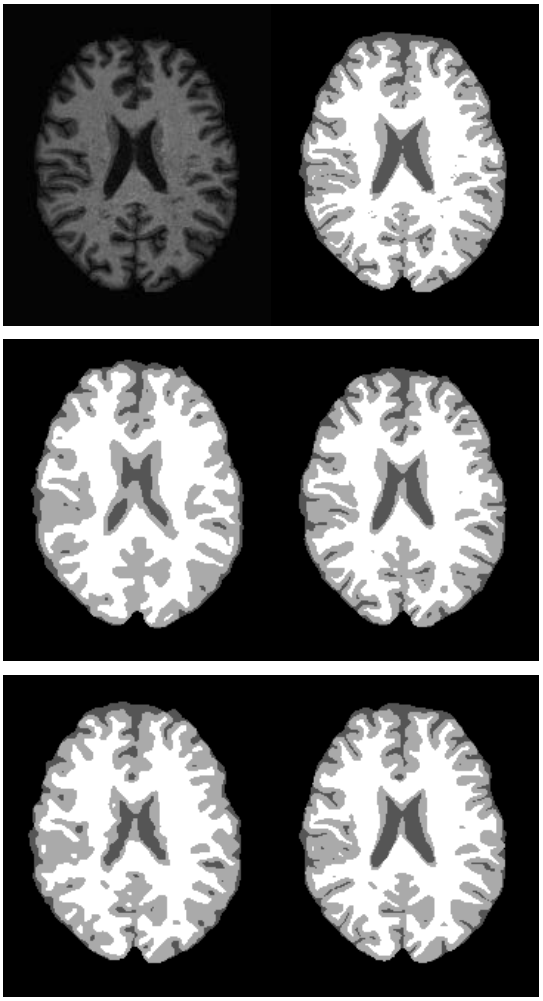


Figure 16: Segmentation of Brain Tissues for HR and simulated LR (downsampled to 4x lower resolution) and prediction from SR-UResNet. From left to right and top to bottom: HR slice, ground truth, segmentation for LR simulated slice (downsampling operator - truncating outer 3D k-space and 4x lower resolution), corresponding segmentation for prediction from SR-UResNet, segmentation for LR simulated slice (downsampling operator - cubic spline interpolation and 4x lower resolution), corresponding segmentation for prediction from SR-UResNet.

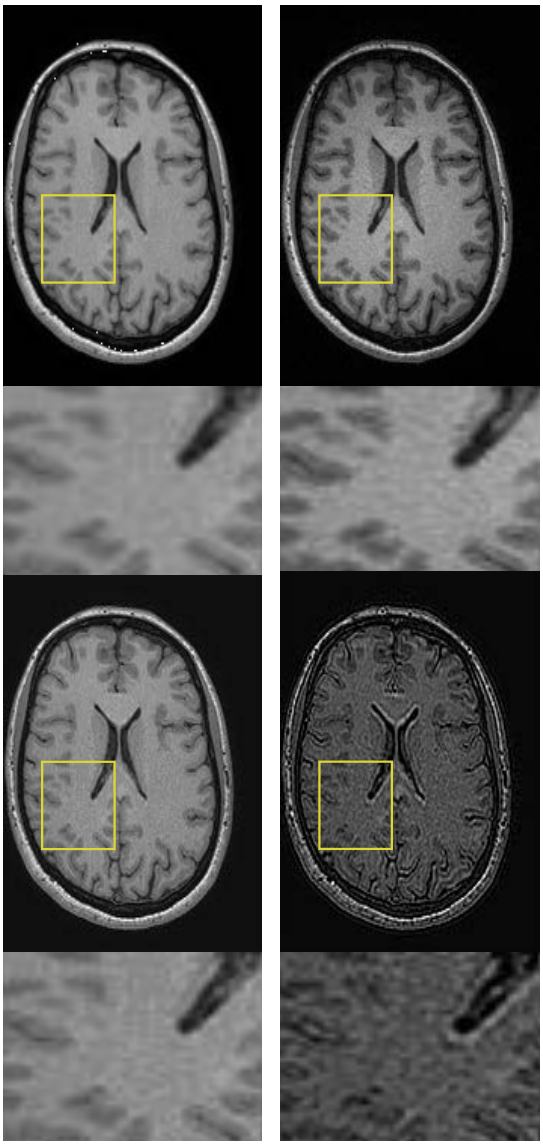


Figure 17: left to right and top to bottom (Input - 1.5T MRI image axial slice, corresponding 3T MRI axial slice, prediction when tested with model trained on 3T MRI as HR and Gaussian Filtered LR with sigma value 0.5, prediction when tested with model trained on 3T MRI as HR and Gaussian Filtered LR with sigma value 1.)

20% for prediction images than LR images obtained by truncating outer 3D k-space for 3 and 4 times lower resolution. And as simulating LR by truncating outer 3D k-space more closely resembles actual LR MRI acquisition, our method looks promising.

In Fig. 18, we have visualized the feature maps of one of the layers after training completed, specifically the last layer before (1x1x1) convolution (Refer: Section 3.3.2 for architecture of SR-UResNet). As there were 32 feature maps in this layer, we show the axial image slices for 6 random feature maps. The (1x1x1) convolution effectively learns to reduce the dimension from this state and reconstructs the image. It is shown in Fig. 18, the features that the model extracted for one of the test volume. These extracted feature are mapped non-linearly and reconstructed. And thus, the three steps for super resolution - feature extraction, non-linear mapping and reconstruction, all solved automatically by the network.

Though we obtained superior results for brain MRI super resolution, however, these results are for the simulated LR. The major challenge is the lack of dataset for actual HR and LR as acquired by scanners. Most of the published works simulate LR from HR by some downsampling operator. In this work, we downsampled with different strategies to validate that super resolution using deep learning approach is better than traditional

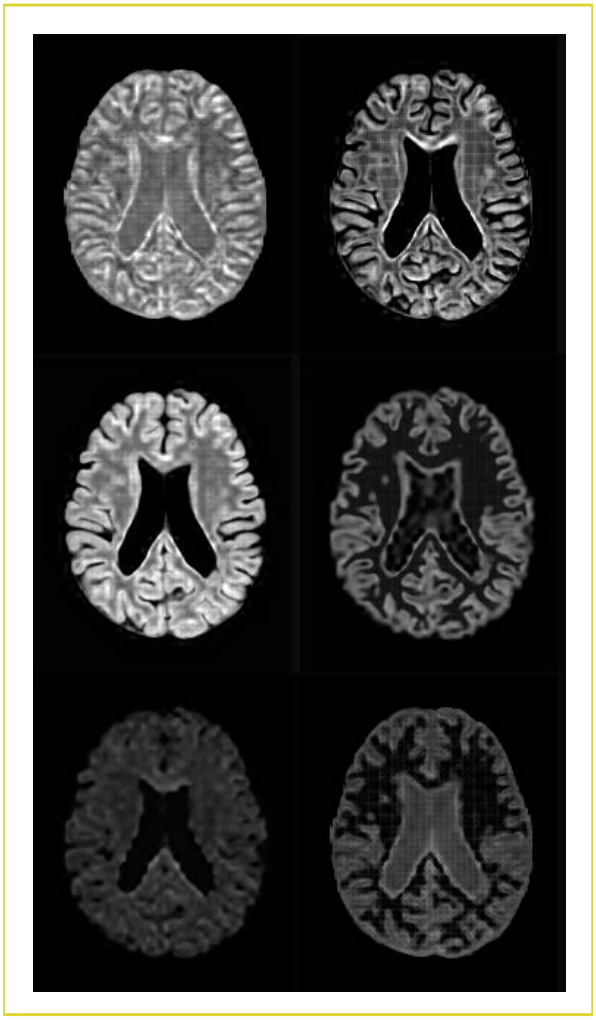


Figure 18: Visualization of the Latent Space. 6 random feature maps for layer before (1x1x1) convolution of SR-UResNet are shown.

methods for diverse degradation problem. The next step forward would be validating the model with actual LR and HR images as acquired by scanners.

6. Conclusions

In this study, two deep learning architecture called 3D-LSRN and SR-UResNet were proposed for super-resolution of brain MRI images. The simulated LR were obtained by gaussian blurring, cubic spline interpolation, and truncating outer 3d k-space. The 3D-LSRN was successfully validated for deblurring the MRI images. The other proposed architecture SR-UResNet was successfully validated for super-resolution by training with downsampled images obtained by cubic spline interpolation and truncating outer 3D k-space. PSNR, SSIM and DSC for segmentation of tissues and subcortical structures were used as objective evaluation criteria and actual predictions were visualized for subjective evaluation. SR-UResNet outperformed traditional bicubic and other CNN method for all downsampling levels and strategies. Thus, our proposed frameworks can be a step towards obtaining super-resolution images from actual LR images which would allow reduction in scan time with same image quality.

7. Acknowledgments

I would like to thank my supervisors Dr. Xavier Llado and Dr. Arnau Oliver for their crucial suggestions and feedback. I am also grateful for MAIA selection committee for selecting me and European Union for funding this studies.

References

Agustsson, E., Timofte, R., 2017. Ntire 2017 challenge on single image super-resolution: Dataset and study, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 126–135.

Anna Nowogrodzki, Nature, 2018. The worlds strongest MRI machines are pushing human imaging to new limits. `https:///www.nature.com/articles/d41586-018-07182-7`. [Online; accessed 2-May-2019].

Bahrami, K., Shi, F., Rekik, I., Shen, D., 2016a. Convolutional neural network for reconstruction of 7t-like images from 3t mri using appearance and anatomical features, in: Deep Learning and Data Labeling for Medical Applications. Springer, pp. 39–47.

Bahrami, K., Shi, F., Zong, X., Shin, H.W., An, H., Shen, D., 2016b. Reconstruction of 7t-like images from 3t mri. IEEE transactions on medical imaging 35, 2085–2097.

Bevilacqua, M., Roumy, A., Guillemot, C., Alberi-Morel, M.L., 2012. Low-complexity single-image super-resolution based on nonnegative neighbor embedding .

Chen, Y., Shi, F., Christodoulou, A.G., Xie, Y., Zhou, Z., Li, D., 2018a. Efficient and accurate mri super-resolution using a generative adversarial network and 3d multi-level densely connected network, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 91–99.

Chen, Y., Xie, Y., Zhou, Z., Shi, F., Christodoulou, A.G., Li, D., 2018b. Brain mri super resolution using 3d deep densely connected neural networks, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE. pp. 739–742.

Clevert, D.A., Unterthiner, T., Hochreiter, S., 2015. Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289 .

Dong, C., Loy, C.C., He, K., Tang, X., 2016a. Image super-resolution using deep convolutional networks. IEEE transactions on pattern analysis and machine intelligence 38, 295–307.

Dong, C., Loy, C.C., Tang, X., 2016b. Accelerating the super-resolution convolutional neural network, in: European conference on computer vision, Springer. pp. 391–407.

Frahm, J., Fransson, P., Krüger, G., 1999. Magnetic resonance imaging of human brain function, in: Modern techniques in neuroscience research. Springer, pp. 1055–1082.

Girshick, R., 2015. Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, pp. 1440–1448.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708.

Irani, D.G.S.B.M., 2009. Super-resolution from a single image, in: Proceedings of the IEEE International Conference on Computer Vision, Kyoto, Japan, pp. 349–356.

Jog, A., Carass, A., Prince, J.L., 2014. Improving magnetic resonance resolution with supervised learning, in: 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 987–990.

Kim, J., Kwon Lee, J., Mu Lee, K., 2016. Accurate image super-resolution using very deep convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1646–1654.

Laurentius Huber, 2018. Locations of 7t MRI worldwide.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al., 2017. Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4681–4690.

Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K., 2017. Enhanced deep residual networks for single image super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 136–144.

Lin, M., Chen, Q., Yan, S., 2013. Network in network. arXiv preprint arXiv:1312.4400 .

Manjón, J.V., Coupé, P., 2016. volbrain: an online mri brain volumetry system. Frontiers in neuroinformatics 10, 30.

Nie, D., Trullo, R., Lian, J., Wang, L., Petitjean, C., Ruan, S., Wang, Q., Shen, D., 2018. Medical image synthesis with deep convolutional adversarial networks. IEEE Transactions on Biomedical Engineering 65, 2720–2730.

Odena, A., Dumoulin, V., Olah, C., 2016. Deconvolution and checkerboard artifacts. Distill URL: `http://distill.pub/2016/deconv-checkerboard`, doi:`10.23915/distill.00003`.

Patenaude, B., Smith, S.M., Kennedy, D.N., Jenkinson, M., 2011. A bayesian model of shape and appearance for subcortical brain segmentation. Neuroimage 56, 907–922.

Pham, C.H., Díez, C.T., Meunier, H., Bednarek, N., Fablet, R., Passat, N., Rousseau, F., 2017. Simultaneous super-resolution and segmentation using a generative adversarial network: Application to neonatal brain mri .

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

Sánchez, I., Vilaplana, V., 2018. Brain mri super-resolution using 3d generative adversarial networks. arXiv preprint arXiv:1812.11440 .

Sanghyun Son, 2017. Pytorch version of the paper 'enhanced deep residual networks for single image super-resolution' (cvprw 2017). `https://github.com/thstkdgus35/EDSR-PyTorch`. [Online; accessed 13-Feb-2019].

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z., 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1874–1883.

Sun, J., Shum, H.Y., 2015. Image super-resolution using gradient profile prior. US Patent 9,064,476.

Tai, Y.W., Liu, S., Brown, M.S., Lin, S., 2010. Super resolution using edge prior and single image detail synthesis, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE. pp. 2400–2407.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., et al., 2004. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing 13, 600–612.

Yang, C.Y., Ma, C., Yang, M.H., 2014. Single-image super-resolution: A benchmark, in: European Conference on Computer Vision, Springer. pp. 372–386.

Yang, J., Wright, J., Huang, T.S., Ma, Y., 2010. Image super-resolution via sparse representation. IEEE transactions on image processing 19, 2861–2873.

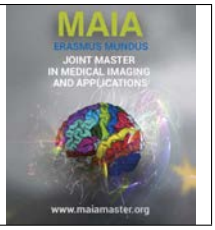
Zeyde, R., Elad, M., Protter, M., 2010. On single image scale-up using sparse-representations, in: International conference on curves and surfaces, Springer. pp. 711–730.

Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. IEEE transactions on medical imaging 20, 45–57.

Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y., 2018. Image super-resolution using very deep residual channel attention networks, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 286–301.

Medical Imaging and Applications

Master Thesis, June 2019



Functional specialization in the ventral visual stream examined with convolutional neural network-derived visual representations

Elizaveta Genke, Katja Seeliger, Marcel van Gerven

Donders Centre for Cognition, Radboud University, Nijmegen, The Netherlands

Abstract

Effective object recognition is one of the most fascinating human brain functions. Neuroimaging studies have discovered separate areas in the visual ventral stream that are responsible for recognizing natural scenes (PPA), faces (FFA) and objects (LOC). It is underexplored why this split is necessary for the function of the visual system, and how it occurs. Recent research has found similarities in convolutional feed-forward network representations and hierarchical representations in the brain. Our objective in this study is to analyse hypothesized brain representations extracted from three convolutional feed-forward neural networks with identical architecture that were trained for face classification (VGG-Face), object classification (Imagenet VGG-16) or scene classification (Places365-VGG16). We used voxel-wise encoding models to map where across the cortex the different convolutional neural network representations match with brain activity when exposed to naturalistic stimuli. The neuroimaging data used in the study is a massive single-subject functional MRI data set where the participant has been exposed to 23 hours of naturalistic video data (a TV series). Using the described methodology we show that models built on different feature sets showed different performance in the ROIs, but the differences were small.

Keywords: functional MRI, brain imaging, convolutional neural networks

1. Introduction

The functioning of the human brain is a mystery that people have been trying to solve for a long time. How is the information from different types of sensors processed in our brain? What are the internal representations that different categories have in the brain? Technological advancement in medical imaging gives us an opportunity to get insights into the human brain and finally answer some of the questions.

The human brain's visual system has been a subject of great interest. The visual system in primates is capable of object recognition under changing circumstances such as: lighting conditions, angle of view and distance. These changing circumstances remain to be a challenge for computer vision systems. Previous studies have shown that the human visual cortex consists of early visual areas V1, V2, V3, V4/V8. The information from these areas is propagated to the inferior temporal cortex (IT) and posterior parietal cortex. Two pathways are present in visual cortex: dorsal stream, also known as the *where*/*how* stream and the ventral or *what* stream (Grill-Spector and Malach, 2004).

There is evidence from neuroimaging studies that there are regions in the ventral visual stream that are selective to specific categories (Kanwisher, 2010). Some areas in the visual ventral stream were found to be responsible for different categories, namely FFA - fusiform face area, PPA - parahippocampal place area and EBA - extrastriate body area. Having these specialized regions may be evolutionarily advantageous and further research is encouraged to discover new specialized regions and to explore how they interact with each other.

There is ongoing research for discovering categorical specialization of the human brain. One of the most remarkable studies was performed by Huth et al. (2012) where the researchers argue that rather than having distinct areas in the brain for each category the human brain has an internal semantic space. Using principal component analysis (PCA) they have identified the first

few semantic dimensions and projected them onto the cortical surface. They have also shown that these dimensions and their cortical organisation are shared between individuals.

As in many other domains deep learning is appearing as a new analysis method in neuroscience. Largely inspired by neuroscience itself, it brings us insight into many complex problems. The operations that take place in deep learning (sum of weighted signals, max pooling, etc.) are mainly influenced by the way neurons work. The convolutional layers in more recent networks were inspired by the organisation of the visual system (Fukushima, 1980).

There is a large body of research in the field of comparing representations learned by convolutional neural networks to the functioning of the brain. Some studies concluded that visual perception is similar in both human brains and in deep neural networks. The studies also concluded that the hierarchical structure holds, so that the learned early layer representations that are more responsive to simple geometric features (like borders, corners, gradients in color) are the most predictive about early visual areas while later layers are more correlated with higher cortex areas.

Our hypothesis is based on the assumption that different input statistics make the networks learn different features of the images. Similarly, different areas in the visual ventral stream are more specialized in getting features related to its role in category recognition. So the internal representations of pre-trained networks and visual stream areas should be similar and the models made from the features derived from different networks trained on a specific category should correlate most with the corresponding functional areas in the visual ventral stream responsible for that category.

2. State of the art

There has been a lot of new research in comparing representations learned by convolutional feed-forward neural networks to sensory systems (mostly the visual system) (van Gerven, 2017; Hassabis et al., 2017; Kietzmann et al., 2018; Kriegeskorte, 2015). Several studies were conducted revealing the connection between the features learnt by convolutional neural networks and brain internal representations (Güçlü and van Gerven, 2015; Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014). The findings of the study by Cichy et al. (2016) is expanding these correlations to the temporal space by analyzing both fMRI and MEG data.

Despite the increasing popularity of the usage of convolutional feed-forward neural networks in the computational neuroscience field, researchers remain sceptical about their explanatory power (Kay, 2018). The downside of models built using deep neural networks is that the networks appear as a ”black box” to scientists due to having a large number of parameters. A better understanding of convolutional neural networks is necessary for the computational neuroscience to benefit more from them.

As an attempt to find the most brain-like network, the Brain-Score platform was created (Schrimpf et al., 2018). This scoring system consists of neural and behavioural benchmarks that represent visual ventral stream neural activity and object recognition. The Brain-Score platform allows any neural network to be evaluated on how similar to the brain it is. Currently, the best performing architectures at the competition are DenseNet-169, CORnet-S, and ResNet-101.

DenseNet-169 is achieving the highest overall performance and has the best results in predicting IT neurons activity. DenseNet-169 is a network architecture in which each layer is connected to the other in a feed-forward manner (Huang et al., 2017). These connections allow the model to re-use learnt features and avoid the problem of vanishing gradient. DenseNet-201 and DenseNet-121 also yield high results on the Brain-Score benchmarks.

CORnet-S is the second best performing model (Kubilius et al., 2018). The CORnet models were biologically inspired to be closer to the brain, e.g. all models have four stages that are associated with visual areas V1, V2, V4 and IT. CORnet-S is a shallow recurrent model that takes advantage of skip connections and mostly inspired by ResNets.

ResNet-101 is the third place network in the Brain-Score benchmark. This network has the best results in predicting behavioural data. ResNets are residual networks that allow for the training of very deep models due to shortcut connections (He et al., 2016). ResNet-152 and ResNet-50 also perform well on the benchmark having the sixth and seventh positions respectively.

We are aiming towards a different form of comparison however. To the best of our knowledge, it is unclear whether models trained for different goals also map onto some known functional specializations for object recognition like places, houses or faces (Kanwisher, 2010). We are using a huge data set of video stimuli with rich semantic content and we are aiming to analyze how networks of identical architectures are able to explain the brain activity in the areas of the visual ventral stream.

Most common tools for comparisons of internal representations in convolutional neural networks and the human brain are encoding models and representational similarity analysis. Here we are using the same voxel-wise encoding model methodology as in the Nishimoto et al. (2011) since it has been proven to work well with the continuous stimuli.

Figure 1: Examples of stimuli frames

3. Material and methods

3.1. Data set

To test the hypothesis we used the *Doctor Who* data set. This data set consists of 23 hours of data of a single participant exposed to video stimuli. A detailed description of the data set can be found in (Seeliger et al., 2018).

3.1.1. fMRI

The data was acquired using Siemens 3T MAGNETOM Prisma with a Siemens32-channel head coil (Siemens, Erlangen, Germany). Functional scans were acquired with a T2* weighted multi-band echo planar imaging pulse sequence (TR=700ms). The volumes have dimensions of $88 \times 88 \times 64$ with a voxel size of 2.4 mm^3. At the end of most of the sessions, a structural scan was acquired with a T1 MP RAGE weighted sequence. Its dimensions are $256 \times 256 \times 192$ with a voxel size of 1 mm^3.

Alignment of the functional scans was made with FSL 5.0 software (Jenkinson et al., 2012). First, all the volumes were aligned with the middle volume in their run. Then the transformations required to align the middle volumes of the different runs to the middle volume of the very first run were calculated and applied to all the volumes in the runs. No other pre-processing was applied.

3.1.2. Stimuli

The training set stimuli that were presented to the subject are 30 episodes of the Dr.Who TV-series from Season 2, 3 and 4 (after re-launch of 2005). Episodes were divided in runs of approximately 12 minutes long. Images of video frames were cropped and resized to have near-equal length and height. A cyan fixation cross was added to the center of the frames. Examples of frames can be found in figure 1.

For test videos Pond Life and Space / Time were used. Pond Life is a mini-series of 5 narrative 1-minute-episodes. Space / Time consists of two mini-episodes of 3 minutes each. These episodes were repeated 22 times.

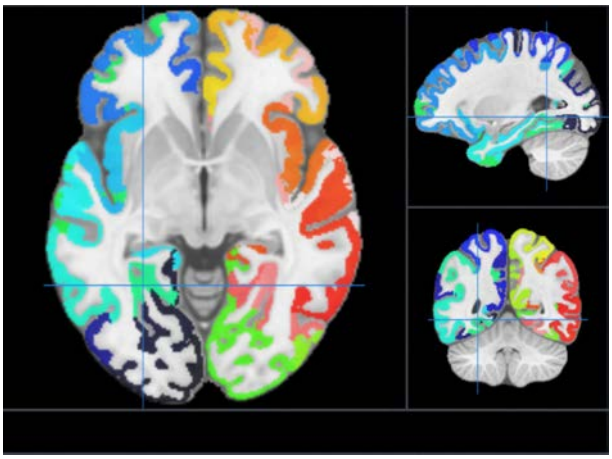


Figure 2: Glasser atlas. The pointer is at the VMV2 area

3.1.3. Localizers

Functional localizers of the visual ventral stream were collected for the subject and estimated as contrasts using FSL. Images of faces, animals, daily objects and human bodies without faces were used. These localizers were verified using *Neurosynth*, Yarkoni et al. (2011) by comparing to other studies.

The localizers include:

- LOC - lateral occipital complex, the area that is activated when images of animals, human bodies without faces and daily objects are shown to the subject.
- FFA - fusiform face area, the cortical region that activates when images of faces are shown.
- OFA - occipital face area, the cortical region that is activated when images of parts of faces are shown.

3.1.4. Glasser atlas ROIs

Another set of ROIs was generated using the Glasser atlas (Glasser et al., 2016). This atlas was built using multi-modal precisely-aligned MR images of 210 adults. It consists of 180 cortical areas. The areas were obtained based on changes in structure, functionality, connectivity, and topography. This atlas is one of the most detailed cortical atlas parcellations available nowadays. It can be found as a surface parcellation for Freesurfer, Mills (2016) and as a volumetric parcellation (Horn, 2016). The volumetric parcellation is in MNI152 2009a non-linear asymmetrical space (Fonov et al., 2009).

To explore ventral visual stream specialization the following cortical areas were chosen:

- FFC (fusiform face complex) is more activated in FACES-AVG (face images vs average of the other three categories: bodies, tools, and places) and FACES-SHAPES (face images vs neutral objects)

contrasts. It is called complex due to internal heterogeneity that could imply that there may be subdivisions.

- PIT complex is also activated in FACES-AVG and FACES-SHAPES contrasts. The border between FFC and PIT complexes is not significant. PIT complex may overlap with OFA.

- VVC (ventral visual complex) and V8 form a heavily myelinated core of the ventral visual stream. They are strongly activated in the PLACE-AVG (place images vs average of the other three categories: bodies, tools, and faces) and the TOOL-AVG (tool images vs average of the other three categories: bodies, faces, and places) contrasts and strongly deactivated in the FACE-AVG contrast.

- VMV1, VMV2, VMV3 are the ventro-medial visual areas that lie between V2, V3 and V4 and parahippocampal areas. They are more deactivated in BODY-AVG (images of bodies vs average of the other three categories: faces, tools, and places) contrast and VMV3 is more activated in TOOL-AVG contrast.

A more detailed description of the areas can be found in Glasser article's supplementary material (Glasser et al., 2016).

The *elastix* toolbox, Klein et al. (2010) was used to register the volumetric version of the atlas into the functional space. First, MNI152 2009a nonlinear asymmetrical volume was registered to the structural scan using non-rigid transformation: affine and b-spline. After that, the rigid transformation was estimated from structural scan to the functional one. The transform parameters files were then used to register the atlas to the functional space.

3.1.5. Neural networks

Three pre-trained convolutional neural networks of VGG16 architecture were chosen for extracting the features from the stimuli images: VGG16 trained on ImageNet, Simonyan and Zisserman (2014), VGG-Face, Parkhi et al. (2015) and VGG-Places365 (Zhou et al., 2017). Three networks with the same architecture were chosen since our goal was to exclude the influence of a specific architecture.

VGG16 is a convolutional neural network with small size filters (3x3) and a depth of 16 layers. It accepts RGB images of the size 224×224. The images are fed to the convolutional layers first. Three fully connected layers are stacked on top of 13 convolutional layers resulting in 138 million parameters to learn. The configuration also includes five max pooling layers that follow some of the convolutional layers. The architecture is shown in the figure 3. It was one of the top performing networks on ILSVRC2014 in localisation and classification tasks. It is one of the most widely used neural networks in the computer vision field but it has been outperformed on the ImageNet competition by more complex networks. Object recognition models from the VGG family have been used to study the similarities of convolutional neural network representations to biological ones (Güçlü and van Gerven, 2015).

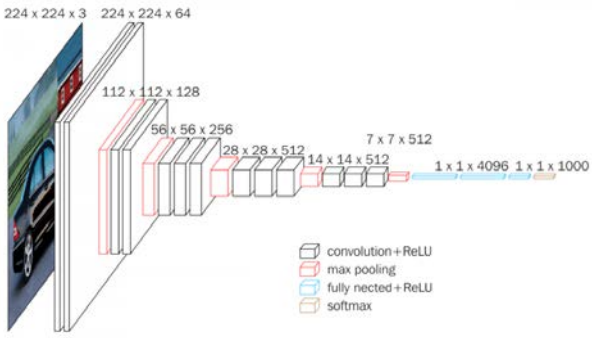


Figure 3: VGG16 architecture (Munneb, 2018)

VGG16 trained on ImageNet is suitable for recognition of daily life objects. ImageNet is the biggest hand-labeled data set available nowadays of about 10 million images with 500-1000 images per category and 1000 categories. It consists of photographs of various objects collected from Flickr and other search engines. Starting from 2010 Large Scale Visual Recognition Challenges on classification, localisation and detection tasks were held (Russakovsky et al., 2015). We expect the features extracted from this network to be mapped to LOC and VMV1-3 areas since those areas were found to be activated when images of tools and objects were presented.

The Places365-VGG16 network is suitable for the recognition of images of places. Places365-VGG16 was trained on the Places365-Standard data set that contains about 1.8 million images in 365 categories with 5000 images per category at most. This network was trained for scene recognition and achieved the highest Top-1 accuracy (about 55.2 %) and the second best result in Top-5 accuracy (about 85 %). The encoding model built on the features extracted with this network is expected to have better performance in VVC and V8.

VGG-Face is the network that classifies images of faces. VGG-Face was trained on the Face data set that consists of 2.6 million images of about 2.6 thousand celebrities (mostly actors). For testing Labeled Faces in the Wild and YouTube Faces data sets were used. It was achieving 100%-EER (Equal Error Rate) of 96.9 on the Labeled Faces in the Wild data set. We expect that the features obtained from the VGG-Face network are going to be highly correlated with the brain activity in FFA, OFA, FFC and PIT ROIs.

VGG-Face, however, has some particularities in preprocessing. The first step is finding a face in the picture. The cropped face image with some offset is then fed into

Table 1: An overview of pre-trained networks that were used for further analysis

Name of the pre-trained network	Tasks	Number of classes	Data set size	Number of images per category	Accuracy
ImageNet VGG-16	object classification and localisation	1000	10 million	500-1000	classification error 0.07405 in classification task in ILSRVC2014
Places365-VGG16	scene recognition	365	1.8 million	up to 5000	Top-5 accuracy of 85.08% on the validation set of Places365 data set
VGG-Face	face recognition	2622	2.6 million	about 1000	100%-EER of 96.9 on the Labeled Faces in the Wild data set

the network. With this pre-processing step, we would not be able to get feature sets from three consecutive frames. Since the model based on features from only one frame had poor performance we have decided to omit the cropping step.

The summarized information about the networks is given in the table 1.

Due to the difference of the input data statistics the networks described above should have different weights on different layers.

All the stimuli frames were inferenced through these networks and activation maps were extracted for max-pooling layers and fully connected layers. For building the encoding models we used only features from fc1 and fc2 layers, because they are the most semantic. For extracting the features, the Matconvnet module and Keras package with a back-end in Tensorflow were used. The computation was done on a single server's GPU.

3.1.6. Encoding model

For analyzing the blood-oxygen-level-dependent (BOLD) signal a linear encoding model was built. Encoding models are a popular tool in the neuroscience field. They attempt to predict the changes in BOLD signal depending on changing stimuli (Naselaris et al., 2011). Encoding models consist of several components. The first is the set of stimuli, which in our case are the frames of the TV-series that were presented to the subject. For every volume of functional MRI data, we consider only the first frame as the stimuli image. The second is the set of features that are the result of the non-linear mapping of the stimuli with the pre-trained neural networks. Here we were using activation maps (features) from different networks that we obtained by inferencing the frames. The third is the ROIs in the brain (we use the collected functional localizers and the Glasser ROIs) and the BOLD signal in these voxels. The last part is the algorithm that connects all three components described above. The encoding model for every individual voxel is:

$$Y = X * W \tag{1}$$

where Y is the BOLD signal, W is a weight matrix and X is the feature data set.

A problem arises with the delay of the hemodynamic response. The stimuli X that produces a BOLD response will occur a few seconds before we see it in Y. When using single images as stimuli there are different techniques that can be applied to solve this issue. One of them would be to show an image for a few seconds and then register the BOLD signal a few seconds later. Then for single images, the encoding model would look like:

$$Y(10sec) = X(5sec) * W \tag{2}$$

But having video stimuli we cannot use this strategy. Therefore we have to use a modified version of this like in the work of Nishimoto et al. (2011). With videos, every BOLD signal in Y will be influenced by multiple X beforehand. To solve that, we just allow the model to take information from multiple signals beforehand. So we stack some amount of X points that precede the actual moment:

$$\begin{bmatrix} . \\ Y_{i-1} \\ Y_i \\ Y_{i+1} \\ . \end{bmatrix} = \begin{bmatrix} . & . & . \\ X_{i-d-1} & X_{i-d} & X_{i-d+1} \\ X_{i-d} & X_{i-d+1} & X_{i-d+2} \\ X_{i-d+1} & X_{i-d+2} & X_{i-d+3} \\ . & . & . \end{bmatrix} * \begin{bmatrix} W_0 \\ W_1 \\ W_2 \end{bmatrix} \tag{3}$$

where d is the delay and each W is a weight matrix of dimensions number of features by 1 for different delay steps. Empirically the delays of 7s, 6.3s, and 5.6 s were chosen (our TR is 0.7 s). With that relatively easy modification, we can build a simple model. To solve the equation 3, the ridge regression (also known as L2-regularised regression or Tikhonov regression) was used in order to avoid overfitting. Usually, such models are solved by finding the pseudoinverse matrix of features and multiplying them with BOLD signal vector.

$$W = X^+ * Y \tag{4}$$

To obtain this matrix, SVD is performed on X matrix and then the pseudoinverse of X is:

$$X^+ = V * \Sigma^{-1} * U^T \tag{5}$$

The Σ^{-1} matrix is obtained by taking every entry in Σ matrix and inverting it (taking 1/S). Very small singular values can interfere with it, resulting in very large

values. For that reason L2-regularised regression introduces a parameter. For every entry S in Σ matrix, there would be the Σ^{-1} matrix entry D and it would be obtained with the formula:

$$D_i = \frac{S_i}{S_i^2 + a^2} \tag{6}$$

And then the weight matrix would be found as:

$$W = V * D * U^T * Y \tag{7}$$

Choosing the best value for a is not a simple task. In real life data, the best strategy to choose a value is to try several ones and choose the best one out of cross-validated results.

For our study 15 a values were tested in a five fold cross-validation. For cross-validation 1% of the training data set was held as a validation set (1158 time points). The best mean result among all folds was then individually chosen for each voxel.

We use this method to get from the features to a prediction of the BOLD signal for every individual voxel on the test set.

The correlation between the predictions and the real signal in the test set is obtained per voxel for each feature set. It is calculated as Pearson correlation.

3.1.7. Cortical maps

In order to visualize the correlations on the cortical surface the *pycortex* tool was chosen (Gao et al., 2015). This tool allows us to build surface visualisations of fMRI data. It can project anatomical and functional data onto the cortical surface and interactively inflate and flatten it. In order to do so a Freesurfer run is needed to create white and gray surfaces.

For an illustration of the ROIs an overlay .svg file was generated using *Inkscape*. The contours were drawn manually for each ROI.

3.1.8. ROI analysis

To get more information about the values of the correlations for the voxels inside the chosen ROIs we have decided to build histograms. They would demonstrate the distribution of the values for different networks inside of the ROIs. For building histograms *numpy.histogram* function was used, with the specified range (-1,1) since those are the only values that correlations may take and with the number of bins equal to 50. The histogram for each ROI with the correlations from the chosen three networks were built for the neural networks fully convolutional layers fc1 and fc2.

In order to quantify the similarity between those histograms and decide which network has better performance, the following strategy was proposed. We look at the correlation of the histograms using the *OpenCV compareHist* function. It calculates the correlation between them on the following formula:

$$d(H_1, H_2) = \frac{\sum_I (H_1(I) - \bar{H_1})(H_2(I) - \bar{H_2})}{\sqrt{\sum_I (H_1(I) - \bar{H_1})^2 \sum_I (H_2(I) - \bar{H_2})^2}} \tag{8}$$

Also, the mean correlation value of the distribution of correlations was extracted for each ROI for each model. In order to find the mean value we had to apply Fisher Z-transformation (due to non-linearity of Pearson correlations).

The last thing that we tried was to threshold correlation values and to compute the mean only for significant voxels ($p < 0.01$, Bonferroni corrected over the total number of gray matter voxels).

4. Results

After extracting features from fc1 and fc2 layers (which are related with the most semantic features) for every frame and building voxel-wise encoding models the cortical surface maps were built. They are presented in the figures 4, 5, 6, 7.

Two maps for each of the sets of features extracted from the fully connected layers are available with two sets of ROIs: localizers generated for the subject as contrasts and the ROIs registered from the Glasser atlas. The correlations projected onto the cortical surface are shown in three colors - ImageNet VGG16 is shown in green colors, VGG-Face is shown in red colors and Places365-VGG16 is shown in blue colors. It is important to mention that the colors do not reflect the neural activity but they show in which voxels the correlation between the encoding model predictions and the actual neural activity is high.

The correlations have mostly mapped across the visual areas and some high correlations can be found in the other areas of the cortex (even in the auditory cortex). The auditory cortex can be predicted since there are correlations between specific audio and specific semantics in natural data. The results of the study by Huth et al. (2012) demonstrate that stimuli of natural movies cause brain activity across the cortex based on semantic space mapping of different categories of objects and actions.

The correlations are lower in the earlier visual areas and are mostly of blue color which may indicate that the features related to place categorization highly correlate with basic image features (color, contrast, simple shapes).

Within visual ventral stream areas, the green color is dominant, which means that ImageNet-VGG16 has the best predictions on the brain activity in this area. Yet the ROIs are not characterized by only one color. The mapping is showing an internal inconsistency of colors.

After that the histograms of the correlation values were built for each ROI for each set of features. Those histograms are shown in the figures 8, 10 and 9. The

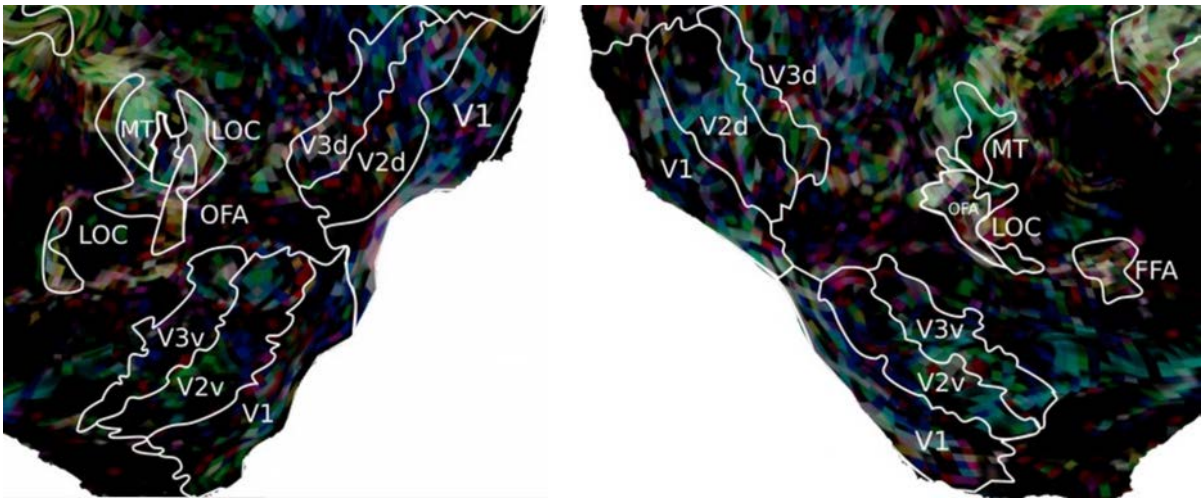


Figure 4: Voxel-wise correlations in RGB channels between measured BOLD responses and BOLD responses predicted based on the 3 feature sets extracted from fc1 layer. The distribution shows where certain feature sets are overrepresented. ImageNet VGG16 in green, VGG-Face in red, Places365-VGG16 in blue channel. Localizers are used as the overlay ROIs

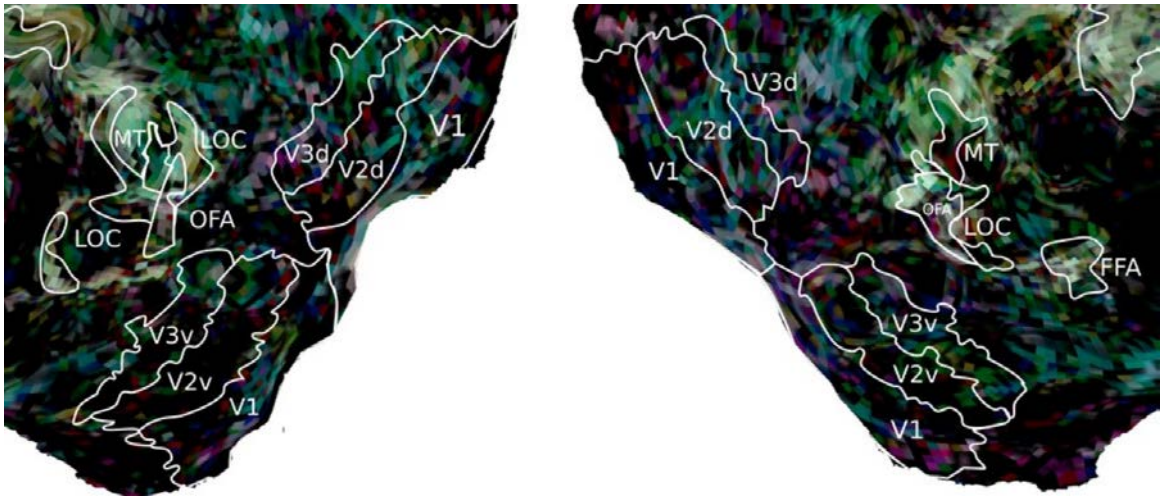


Figure 5: Voxel-wise correlations in RGB channels between measured BOLD responses and BOLD responses predicted based on the 3 feature sets extracted from fc2 layer. The distribution shows where certain feature sets are overrepresented. ImageNet VGG16 in green, VGG-Face in red, Places365-VGG16 in blue channel. Localizers are used as the overlay ROIs

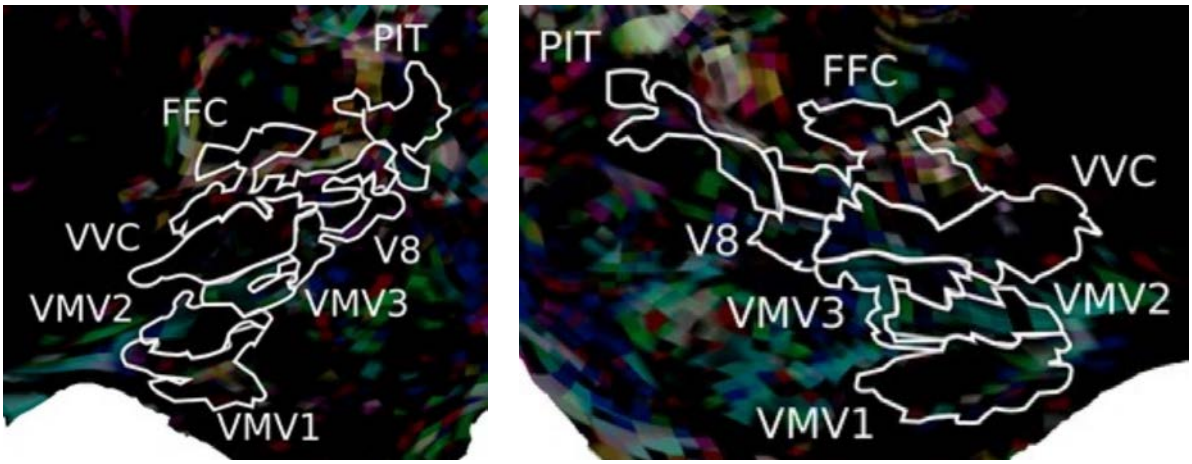


Figure 6: Voxel-wise correlations in RGB channels between measured BOLD responses and BOLD responses predicted based on the 3 feature sets extracted from fc1 layer. The distribution shows where certain feature sets are overrepresented. ImageNet VGG16 in green, VGG-Face in red, Places365-VGG16 in blue channel. Glasser atlas ROIs are used

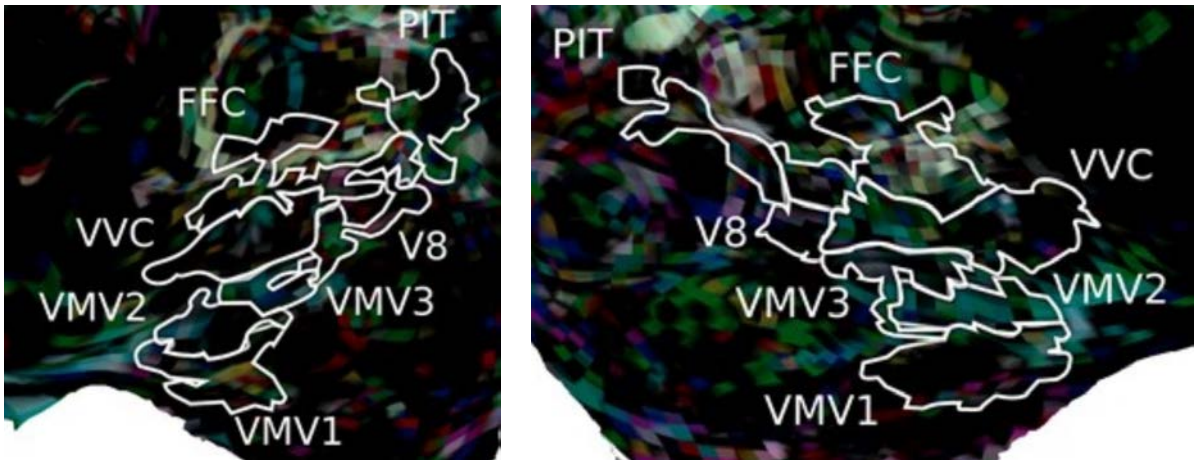


Figure 7: Voxel-wise correlations in RGB channels between measured BOLD responses and BOLD responses predicted based on the 3 feature sets extracted from fc2 layer. The distribution shows where certain feature sets are overrepresented. ImageNet VGG16 in green, VGG-Face in red, Places365-VGG16 in blue channel. Glasser atlas ROIs are used

mean values for each model performance for each ROI were computed and the results are shown in the table 2. They have also been thresholded with 0.19 for analyzing only the voxels where the correlation was found significant. Again the mean value for each model for each ROI was computed and the results are shown in the table 3.

The mean values of the correlations are in accordance with our hypothesis. Face related ROIs (OFA, FFA, FFC, PIT) have higher mean values of correlation with the model built on the features derived from VGG-Face. Object related ROIs (VMV1-3 and LOC) have higher mean values of correlation with the model built on the features derived from ImageNet VGG16. Place related ROIs (VVC and V8) have higher mean values of correlation with the model built on the features derived from Places365-VGG16.

In general, the performance of the models built on features derived from fc2 was worse than the performance of the models built on fc1 features. The highest layer (the one closest to the final category layer) was often not found to be predictive when comparing DNNs and visual system responses. A possible reason is that the artificial set of categories used for model training is not biologically plausible.

The difference between the values is very small however. The histograms of values show close to normal distribution behaviour. This means that many voxels could not be predicted. The correlation of the histograms was calculated for each pair of histograms and those values are presented in the table 4. The calculated values of correlations of the histograms are very high and that means that all models have very similar performance on the ROI level.

Table 2: Mean value of correlations of encoding models within ROIs

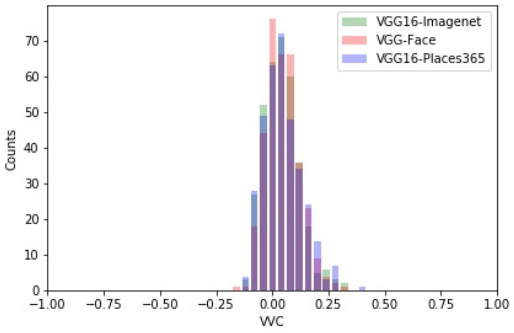
Expected highest correlation set of features	Name of the ROI	ImageNet-VGG16		VGG-Face		Places365-VGG16	
		fc1	fc2	fc1	fc2	fc1	fc2
ImageNet-VGG16	LOC	**0.217**	0.175	0.209	0.171	0.181	0.15
	VMV1	0.079	0.049	**0.086**	0.061	0.082	0.058
	VMV2	**0.123**	0.078	0.061	0.031	0.121	0.092
	VMV3	**0.154**	0.117	0.093	0.064	0.153	0.125
VGG-Face	OFA	0.252	0.205	**0.266**	0.221	0.229	0.19
	FFA	0.157	0.118	**0.174**	0.138	0.125	0.09
	FFC	0.061	0.037	**0.089**	0.062	0.073	0.054
	PIT	0.061	0.043	**0.081**	0.056	0.061	0.065
Places365-VGG16	VVC	0.06	0.039	**0.066**	0.046	**0.066**	0.047
	V8	0.08	0.06	0.068	0.047	**0.096**	0.079

Table 3: Mean value of correlations of encoding models of significant voxels within ROIs. No significant values were found for the model built on VGG-Face features in VMV2

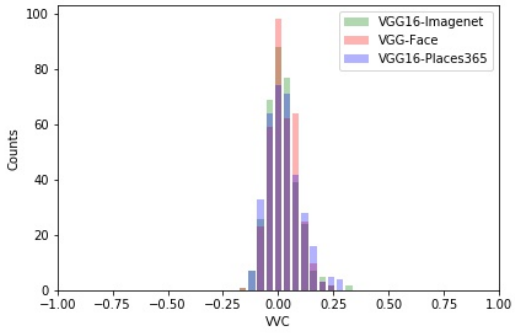
Expected highest correlation set of features	Name of the ROI	ImageNet-VGG16		VGG-Face		Places365-VGG16	
		fc1	fc2	fc1	fc2	fc1	fc2
ImageNet-VGG16	LOC	**0.339**	0.313	0.323	0.295	0.31	0.285
	VMV1	**0.238**	0.223	0.219	0.23	0.237	0.221
	VMV2	**0.236**	0.229	-	-	0.225	0.225
	VMV3	0.25	0.25	0.25	0.237	0.253	**0.255**
VGG-Face	OFA	**0.379**	0.335	0.359	0.326	0.345	0.319
	FFA	0.319	0.285	**0.322**	0.296	0.292	0.293
	FFC	**0.279**	0.264	0.278	0.262	0.261	0.256
	PIT	0.241	0.244	**0.246**	0.236	0.241	0.238
Places365-VGG16	VVC	0.243	0.244	0.228	0.22	0.245	**0.247**
	V8	0.221	0.214	0.214	0.209	0.227	**0.228**

Table 4: Correlations of histograms of correlation values of the networks inside ROIs for two layers

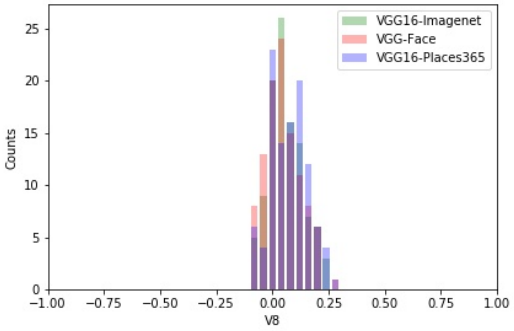
Name of the ROI	Correlation of ImageNet VGG16 with Places365-VGG16		Correlation of ImageNet VGG16 with VGG-Face		Correlation of Places365-VGG16 with VGG-Face	
	fc1	fc2	fc1	fc2	fc1	fc2
LOC	0.953	0.969	0.967	0.986	0.969	0.958
VMV1	0.968	0.976	0.941	0.959	0.975	0.984
VMV2	0.976	0.931	0.806	0.879	0.751	0.7
VMV3	0.944	0.947	0.763	0.83	0.767	0.773
OFA	0.86	0.947	0.861	0.947	0.936	0.94
FFA	0.913	0.913	0.962	0.94	0.913	0.941
FFC	0.98	0.96	0.955	0.957	0.966	0.989
PIT	0.937	0.93	0.935	0.939	0.97	0.913
VVC	0.99	0.99	0.988	0.971	0.979	0.967
V8	0.917	0.913	0.983	0.976	0.891	0.914



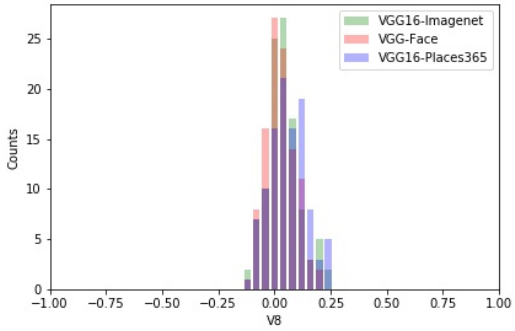
Correlations from fc1 layer in VVC



Correlations from fc2 layer in VVC

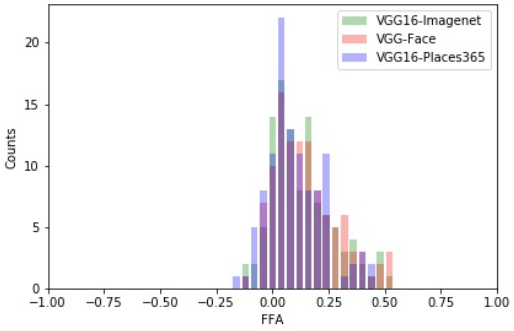


Correlations from fc1 layer in V8

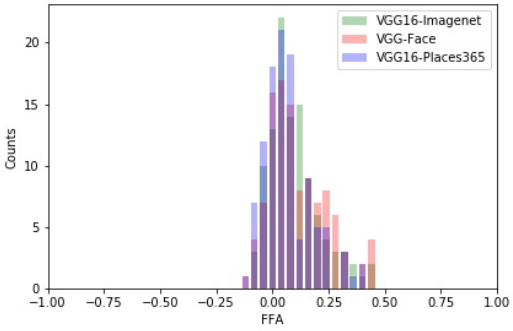


Correlations from fc2 layer in V8

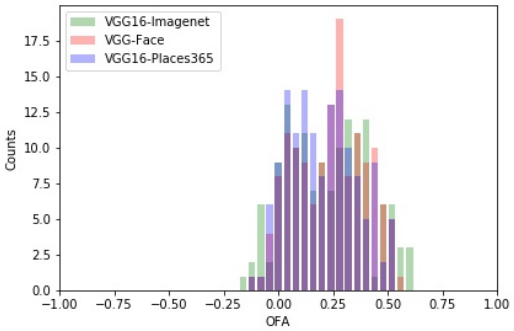
Figure 8: Histograms for places-related ROIs



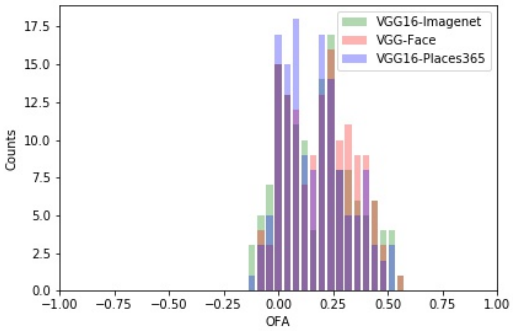
Correlations from fc1 layer in FFA



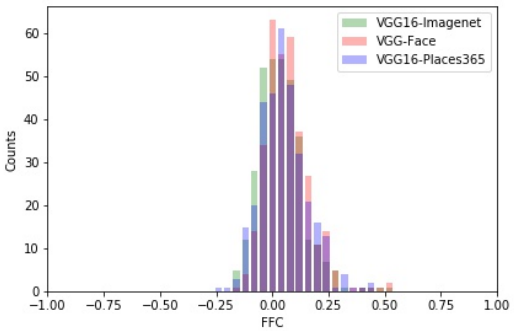
Correlations from fc2 layer in FFA



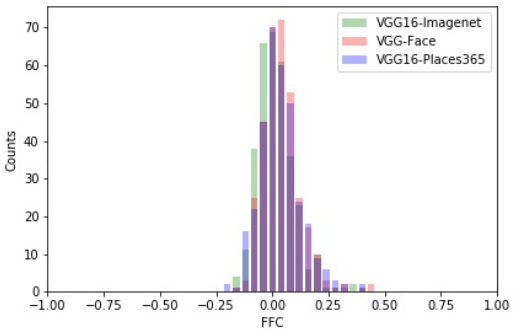
Correlations from fc1 layer in OFA



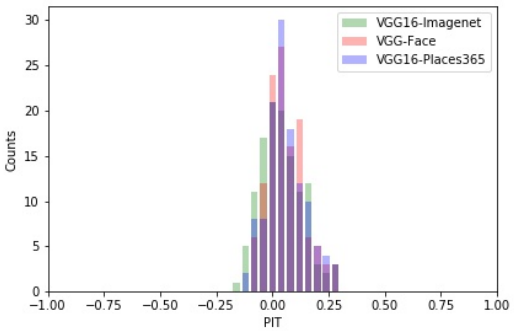
Correlations from fc2 layer in OFA



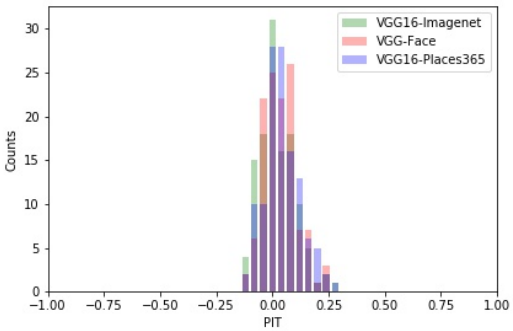
Correlations from fc1 layer in FFC



Correlations from fc2 layer in FFC

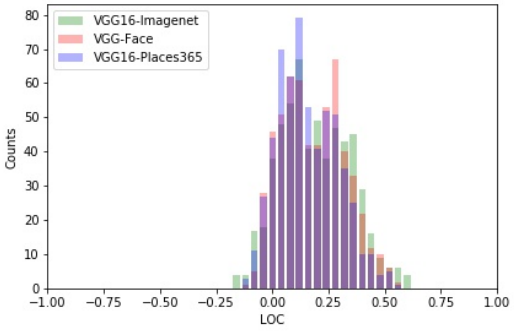


Correlations from fc1 layer in PIT

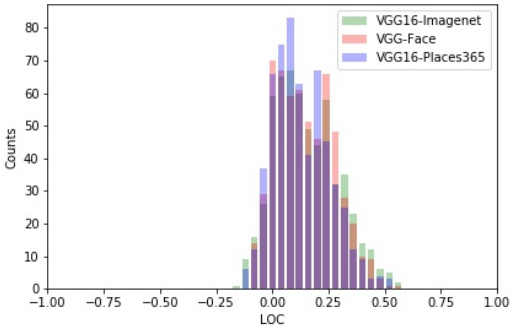


Correlations from fc2 layer in PIT

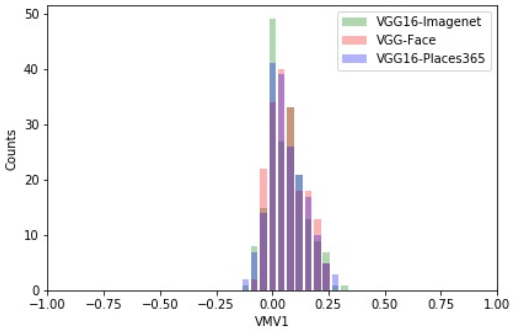
Figure 9: Histograms for face-related ROIs



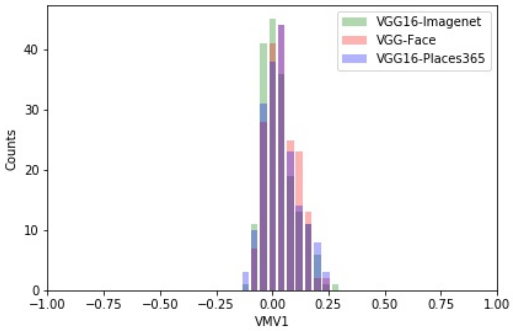
Correlations from fc1 layer in LOC



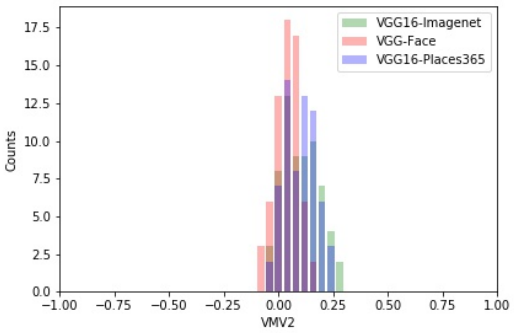
Correlations from fc2 layer in LOC



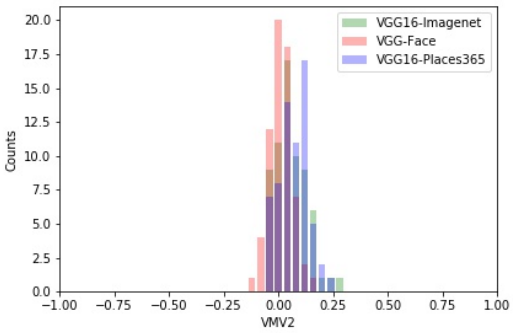
Correlations from fc1 layer in VMV1



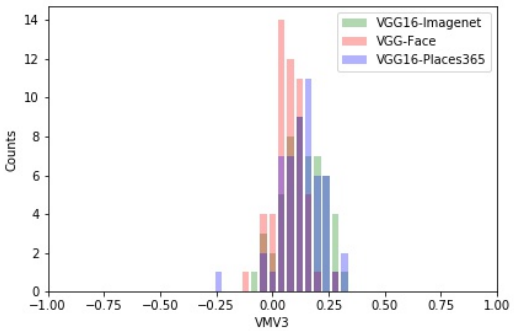
Correlations from fc2 layer in VMV1



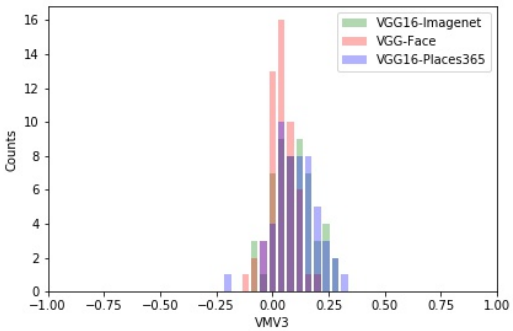
Correlations from fc1 layer in VMV2



Correlations from fc2 layer in VMV2



Correlations from fc1 layer in VMV3



Correlations from fc2 layer in VMV3

Figure 10: Histograms for object-related ROIs

5. Discussion

This work hypothesized that different training of the artificial neural network should result in a split of mapping of features into different areas of the visual ventral stream. Even though means are matching the hypothesis; a closer look at the the differences between the correlation distributions revealed that most of these differences were not significant (Wilcoxon test, p¡0.05).

One of the possible reasons might be low SNR ratio. Since the stimuli was a video signal with a frequency of about 24Hz the BOLD signal may not change much (Poldrack et al., 2011). The encoding model was built using just three time points. Possibly that was not enough to model the BOLD signal change. The value for the delay and *a* parameters in L2-regularized regression were chosen empirically and that also may have contributed.

Another limitation of the encoding model is that it asummes that BOLD response in voxels has a linear time-invariant system behaviour. This is true if the stimuli are static, but it is not guaranteed with dynamic stimuli. I suppose that including motion related features like in the work of Nishimoto et al. (2011) would help to overcome this issue and model the dynamics of brain activity better.

Another issue that may have caused the poor performance of the voxel-wise encoding models is the choice of the architecture. While VGG16 is a popular choice for computer vision problems, more interesting network architectures with plausible biological brain-like properties exist. For example recurrent neural networks and LSTM could produce better results.

Another issue that we have encountered is that the networks assume the presence of an object, place or face. While objects and places were present in almost every frame, this was not true for faces. Also, an ambiguous situation arose when multiple faces were present.

At first, our decision was to zero out the response where there was no face found. In the frames where there were multiple faces found, the closest face to the fixation cross was chosen to be fed into the network. However, this did not work well with our encoding model since it requires three consecutive frames. In the end, it was decided to extract features from frames without pre-processing steps (namely face detection and cropping).

Another detail is that while the networks have very similar architecture, the number of classes they were trained to differentiate is very different for all of them. That may have resulted in different neural capacities of neurons. Also the input statistics were definitely different. It was discussed for ImageNet in Mehrer et al. (2017) but the same goes for VGG-Face which was trained on the images of celebrities and therefore could have some biases that are different from the ones the participant has.

Summarizing the arguments discussed above there is a need for further exploration of how the architecture and input statistics may have changed the performance of models.

Apart from that, the results show that the mean value of the correlations in the recorded functional localizers are higher comparing to the ROIs obtained from Glasser atlas. That may be caused by misregistration of the Glasser atlas. In the article of Glasser et al. (2016) they mention that volumetric registration was not possible due to distortion of the areas and fine parcellation was only possible after multimodal surface-based analysis. It is possible that using their classifier would be a better option for registration. However, this instrument is not yet available and most likely would need multi-modal scans like in HCP project.

We have not explored if the split between network-derived features is more evident for convolutional layers due to the time constraints yet it is an interesting topic to study. Also, the split may be more evident on the last layer as it is shown in the work by (Cichy et al., 2016). However, due to a different number of labels, it is not clear whether that would be a fair comparison.

6. Conclusions

To sum up, further investigation of the ventral stream specialization with the encoding models is necessary. Although the results of our study are not plausible (possibly due to model limitations), we see a potential in the studies that explore how different deep neural networks can explain brain activity. We believe this would be beneficial both for the computational neuroscience field and for artificial intelligence development.

7. Acknowledgments

First of all, I would like to thank my supervisors, Katja Seeliger and professor Marcel van Gerven for their assistance with the thesis and help with all the paperwork and visa issues. I enjoyed working with you on this project and learning new concepts of neuroscience.

I would also like to thank my professors from the University of Burgundy, the University of Cassino and Southern Lazio and the University of Girona. I am thankful for the knowledge you gave me during the courses, for the guidance that you provided for my projects and for the inspiration you gave me to follow a career in the medical imaging field.

This master program was an incredible experience and I am grateful to be chosen for this Erasmus Mundus Joint Master Degree scholarship. Not only has it given me an opportunity to study the fascinating subject but I also got to experience the culture of different European countries and meet friends from all over the world.

I would like to wish luck to all MAIA students in their search for a PhD or job position and thank my Coco Caline family and Soviets. This master would not be the same without you and I hope that our friendship will survive years and distance.

I would like to thank my family for the support and belief in me. This master wouldn't be possible without your encouragement. I would like to also thank Berwout for always being there for me in the moments when I needed your support the most.

References

Cichy, R.M., Khosla, A., Pantazis, D., Torralba, A., Oliva, A., 2016. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Scientific reports 6, 27755.

Fonov, V.S., Evans, A.C., McKinstry, R.C., Almli, C., Collins, D., 2009. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. NeuroImage , S102.

Fukushima, K., 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological cybernetics 36, 193–202.

Gao, J.S., Huth, A.G., Lescroart, M.D., Gallant, J.L., 2015. Pycortex: an interactive surface visualizer for fMRI. Frontiers in Neuroinformatics 9, 23. URL: https://www.frontiersin.org/article/10.3389/fninf.2015.00023, doi:10.3389/fninf.2015.00023.

van Gerven, M.A., 2017. A primer on encoding models in sensory neuroscience. Journal of Mathematical Psychology 76, 172–183.

Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C.F., Jenkinson, M., et al., 2016. A multi-modal parcellation of human cerebral cortex. Nature 536, 171.

Grill-Spector, K., Malach, R., 2004. The human visual cortex. Annu. Rev. Neurosci. 27, 649–677.

Güçlü, U., van Gerven, M.A., 2015. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. Journal of Neuroscience 35, 10005–10014.

Hassabis, D., Kumaran, D., Summerfield, C., Botvinick, M., 2017. Neuroscience-inspired artificial intelligence. Neuron 95, 245–258.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Horn, A., 2016. HCP-MMP1.0 projected on MNI2009a GM (volumetric) in NIfTI format doi:10.6084/m9.figshare.3501911.v5. https://figshare.com/articles/HCP-MMP1_0_projected_on_MNI2009a_GM_volumetric_in_NIfTI_format/3501911.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708.

Huth, A.G., Nishimoto, S., Vu, A.T., Gallant, J.L., 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76, 1210–1224.

Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., M, S.S., 2012. FSL. Neuroimage 62, 782–790.

Kanwisher, N., 2010. Functional specificity in the human brain: a window into the functional architecture of the mind. Proceedings of the National Academy of Sciences 107, 11163–11170.

Kay, K.N., 2018. Principles for models of neural information processing. Neuroimage 180, 101–109.

Khaligh-Razavi, S.M., Kriegeskorte, N., 2014. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS computational biology 10, e1003915.

Kietzmann, T.C., McClure, P., Kriegeskorte, N., 2018. Deep neural networks in computational neuroscience. bioRxiv , 133504.

Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.W., 2010. elastix: A toolbox for intensity-based medical image registration. IEEE Transactions on Medical Imaging 29, 196–205. doi:10.1109/TMI.2009.2035616.

Kriegeskorte, N., 2015. Deep neural networks: a new framework for modeling biological vision and brain information processing. Annual review of vision science 1, 417–446.

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D.L., DiCarlo, J.J., 2018. CORnet: Modeling the neural mechanisms of core object recognition. BioRxiv , 408385.

Mehrer, J., Kietzmann, T.C., Kriegeskorte, N., 2017. Deep neural networks trained on ecologically relevant categories better explain human IT, in: Poster presented at Conference on Cognitive Computational Neuroscience, Submission ID.

Mills, K., 2016. HCP-MMP1.0 projected on fsaverage doi:10.6084/m9.figshare.3498446.v2. https://figshare.com/articles/HCP-MMP1_0_projected_on_fsaverage/3498446.

Munneb, H., 2018. VGG16 convolutional network for classification and detection https://neurohive.io/en/popular-networks/vgg16/.

Naselaris, T., Kay, K.N., Nishimoto, S., Gallant, J.L., 2011. Encoding and decoding in fMRI. Neuroimage 56, 400–410.

Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., Gallant, J.L., 2011. Reconstructing visual experiences from brain activity evoked by natural movies. Current Biology 21, 1641–1646.

Parkhi, O.M., Vedaldi, A., Zisserman, A., 2015. Deep face recognition, in: British Machine Vision Conference.

Poldrack, R.A., Mumford, J.A., Nichols, T.E., 2011. Handbook of functional MRI data analysis. Cambridge University Press.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) 115, 211–252. doi:10.1007/s11263-015-0816-y.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N.J., Rajalingham, R., Issa, E.B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., et al., 2018. Brain-score: which artificial neural network for object recognition is most brain-like? BioRxiv , 407007.

Seeliger, K., Sommers, R.P., Güçlü, U., Bosch, S.E., van Gerven, M.A.J., 2018. A large single-participant fMRI dataset for probing brain responses to naturalistic stimuli in space and time. Dataset published on Donders Data Repository http://11633/di.dcc.DSC_2018.00082_134.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556.

Yamins, D.L., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., DiCarlo, J.J., 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences 111, 8619–8624.

Yarkoni, T., Poldrack, R.A., Nichols, T.E., Van Essen, D.C., Wager, T.D., 2011. Large-scale automated synthesis of human functional neuroimaging data. Nature methods 8, 665.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A., 2017. Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence .

Medical Imaging and Applications

Master Thesis, June 2019

Detection, Segmentation, and 3D Pose Estimation of Surgical Tools Using Deep Convolutional Neural Networks and Algebraic Geometry

Md. Kamrul Hasan, Lilian Calvet, Simone Gasparini, Adrien Bartoli

EnCoV-TGI, Institut Pascal, Clermont-Ferrand, France

Abstract

Surgical instrument detection, segmentation, and pose estimation are crucial computer vision components in Computer-Assisted Intervention (CAI). We propose a new Deep Convolutional Neural Network (DCNN)-based automatic tool segmentation and 3D pose estimation method put at the service of augmented reality (AR) for gesture guidance in laparosurgery. Tool segmentation can be used to increase the depth perception by making the surgical tools visible which were previously occluded by the virtual elements displayed on an augmented laparoscopic video. The 3D pose estimation can be used as constraints to the registration problem between a pre-operative 3D model of an organ and the laparoscopic image, whose solution is at the core of AR in laparosurgery. Although significant research has been done in the recent years, concurrent detection, segmentation, and geometric feature extraction dedicated to 3D pose estimation of the surgical instruments are still very challenging and form an open problem. In this thesis, we present a Single Input Multiple Output (SIMO) DCNN network named ART-Net (Augmented Reality Tool Network) consisting of an encoder-decoder architecture designed to obtain the surgical tool detection, segmentation, and geometric features concurrently in an end-to-end fashion. One of the key ideas to obtain a more accurate segmentation and instrument boundary results than the state-of-the-art is the use of a Full resolution Feature Map Generator (FrG) at the very end of the decoder. We evaluate the proposed segmentation sub-network and compare it against the Fully Convolutional Network (FCN) and U-Net using the publicly available EndoVis (robotic) dataset. The whole proposed network has been further evaluated using the combined EndoVis (non-robotic) and our annotated dataset. The 3D pose of a detected surgical tool is computed using a geometric solver. Tool boundaries along with a point located at the boundary between the tool body and tool head are used as the features for the solver. The obtained mean Intersection over Union (mIoU) and mean Balanced Accuracy (mBA) for the segmentation on EndoVis (robotic) are 81.0% and 93.4% respectively. The proposed ART-Net outperforms both FCN and U-Net respectively by 4.5% and 2.9% in mIoU. The mIoU and mBA of the segmented mask from our network are 88.2% and 97.1% respectively on the combined EndoVis (non-robotic) and our annotated datasets. The mean Arc Length (mAL) error for edge-line (tool boundary) and mid-line are 2.45° and 2.23° respectively whereas the mean euclidean distance error for the tip-point is 9.3 pixels. The average precision, average accuracy, and AUC for the instrument detection are 100.0%, 100.0%, and 1.0 respectively. Our approach outperforms other methods for detection, segmentation, and 3D pose estimation of surgical instruments and is able to solve the ambiguity in registration for AR in laparosurgery. Additionally, the segmented instrument mask is robust and accurate for occlusive surgical instrument visualization in AR. Our annotated dataset and trained model along with the source codes will be made publicly available[1].

Keywords: Computer-Assisted Intervention (CAI), Augmented Reality (AR), Laparosurgery, Deep Learning, Segmentation, 3D pose, Algebraic Geometry.

1. Introduction

Minimally Invasive Surgery (MIS) is a preferable approach for many surgical procedures which reduces operative trauma, blood loss, infection rates, hospitalization and increases speeds of recovery (Bodenstedt et al., 2018; Jaffray, 2005). Although MIS confers

considerable advantages for the patient over open surgery it imposes arduous challenges on surgeon's performance due to issues associated with the Field of View (FoV), hand-eye dis-alignment, and disorientation (Chen et al., 2017).

Augmented Reality (AR) is the fusion of real-world information with computer-generated information. It overcomes the above limitations on the surgeon's performance by overlaying additional information with the real scene through augmentation of the target surgical locations, annotations, labels, 3D reconstruction of organ's inner anatomic structures (Bourdel et al., 2017; Kim et al., 2012). AR can provide better depth perception, interactive visualization and increase surgeon's visual awareness of high-risk surgical targets by accurately overlaying a pre-operative 3D model onto the intra-operative laparoscopic video (Puerto-Souza and Mariottini, 2013). It can be implemented without additional hardware where registration of the pre-operative volume onto the laparoscope's image and occluded object visualization is the arduous difficulties (Ozgur et al., 2017). It is also very difficult to maintain a stable and prolonged overlay of the pre-operative 3D model onto intra-operative video due to having sudden, frequent, and extended occlusions or organ deformations (Puerto-Souza and Mariottini, 2013).

In AR, the extraction of 2D information from 2D laparoscopic image frames to estimate the 3D pose of the surgical instrument has the ability to solve current ambiguities in registration (Cano et al., 2008). The resulting 3D pose information can be used to augment the surgeon's view by projecting artificial 3D cues onto the 2D display which provides additional support by improving depth perception for the surgeon (Wengert et al., 2008). Overlaying a pre-operative 3D model onto the intra-operative laparoscopic image is a strenuous task and requires understanding the spatial relationships between the surgical camera, instruments along with patient anatomy (Pakhomov et al., 2017). A critical component of this overlaying process is segmentation of the surgical instrument which is used to prevent rendered overlays from occluding the instruments (Pezzementi et al., 2009). So, in MIS, robust and automatic surgical instrument segmentation, and 3D pose estimation are essential parts to solve the current limitations of AR applications and for interactive visualizations.

The segmentation of the surgical tool is the most strenuous task due to the presence of smoke, blood, partial occlusions, shadows, specular reflections, motion blur, lighting conditions, scale effect, cauterization, clip, gauze, and complex background textures (Attia et al., 2017; Garcia-Peraza-Herrera et al., 2017b; Pakhomov et al., 2017) as shown in Fig. 1. Agustinos and Voros (2015) used color and shape information along with post-processing and contour detection. Allan et al. (2013) has proposed four colors based Random Forest (RF) (Breiman, 2001) for the classification of the surgical instrument pixels from background tissue pixels. There is a higher possibility of getting coarse segmentation (Pakhomov et al., 2017) of the surgical instruments from those methods. Doignon et al. (2005) developed a segmentation technique based on a discriminant color feature and also designed an adaptive region growing with automatic region seed detection. But, a region growing algorithm depends on the initial seed (Shan et al., 2008) and robust seed selection is often impossible due to the presence of diverse noises in laparoscopic images. Nowadays, semantic segmentation using Convolution Neural Network (CNN) can be a good choice that can provide robust solutions for the surgical instruments segmentation by pixel-wise classifying and assigning their corresponding labels (Ronneberger et al., 2015; Shelhamer et al., 2017). Recently, Garcia-Peraza-Herrera et al. (2017b) applied Fully Convolutional Networks (FCN8s) for binary instrument segmentation. Shvets et al. (2018) used four different encoder-decoder networks which are U-Net (Ronneberger et al., 2015), two modified TernausNet (Iglovikov and Shvets, 2018), and one modified LinkNet (Chaurasia and Culurciello, 2017). Attia et al. (2017); Garcia-Peraza-Herrera et al. (2017a); Pakhomov et al. (2017) used ResNet-101, ToolNetMS/ ToolNetH and hybrid CNN-RNN networks respectively. However, automatic, robust, and accurate segmentation of surgical instruments are highly desirable requirements for interactive AR in MIS even in the presence of the above mentioned different types of noise.

For the overlaying to achieve registration, the position of both the computer generated and real-world information need to be in a common coordinate system. From a practical point of view, the direct use of laparoscopic images for tracking instrument in MIS is better than using extra tracking devices (Feuerstein et al., 2007). In recent years, computer vision algorithms are able to compute the position from the video frame of the camera directly (Davison et al., 2007). To get the 3D position, those algorithms have to select features (landmarks, shades, silhouettes, etc.) directly out of the images and analyze them (Salah et al., 2011; Wang et al., 2012). Feature detection is a computationally expensive and sometimes noisy process. Bourdel et al. (2017); Du et al. (2015); Kim et al. (2012) used the traditional 2D feature-based tracking algorithms which provide poor quality visual guidance due to fall out of the FoV. Recently, Simultaneous Location and Mapping (SLAM) has the potentiality to overcome the previous

[1] https://github.com/kamruleee51/3D-Pose-Estimation-and-Segmentation-for-AR-in-MIS

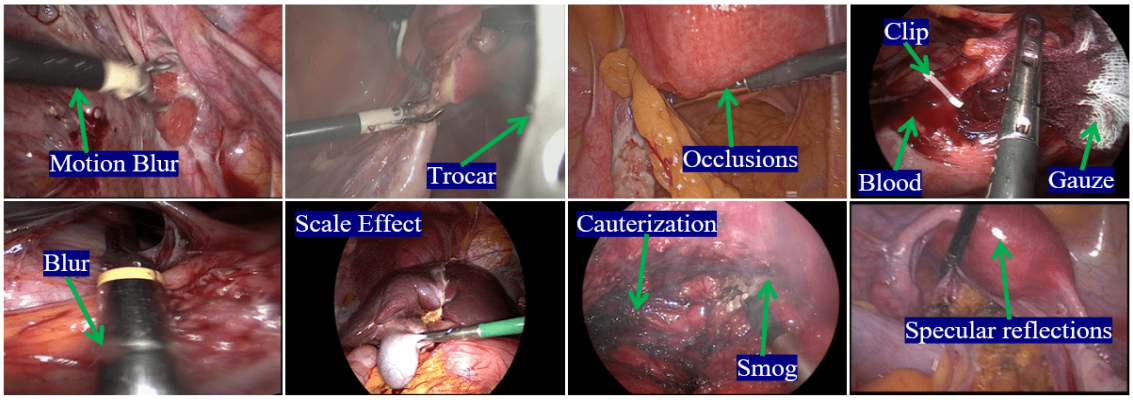


Figure 1: Types of different artifacts present in laparoscopic image. The detection, segmentation, and feature extraction results from ART-Net on those images are shown in Appendix A.

limitations by building an entire 3D map of the internal cavity but is often not robust enough when dealing with tissue deformations and scene illuminations (Artal et al., 2015). Jayarathne et al. (2013); Pratt et al. (2015) have extracted 2D locations where pose estimation is considered as Perspective-n-Point (PnP) (Wu and Hu, 2006) problem. However, pose estimation of a planar marker may provide two ambiguous solutions where the incorrect one should be eliminated (Collins and Bartoli, 2014).

In this thesis, we propose ART-Net for segmenting the laparoscopic images to get surgical instrument's mask and extracting the geometric features simultaneously. The 3D pose of the surgical instrument is a function of instrument physical properties, namely diameter and head length, which are very diverse for clinical applications. To make a generic network which should be irrespective of tool's physical properties, we divide the 3D pose estimation of the surgical instrument into two parts: geometric features extraction from responsible sub-networks of ART-Net and algebraic geometric solution. Finally, the accurate 3D instrument pose is estimated by a geometric solver using the features from the proposed ART-Net and known instrument's physical properties. The 3D pose from our approach is then used for solving the current ambiguities of AR applications in MIS. Additionally, we add detection sub-network to our proposed ART-Net to detect the tool present in the input laparoscopic image frame which helps to decide for estimating pose or not. The proposed network with different responsible sub-networks has been trained in an end-to-end fashion and each sub-network of the ART-Net has different loss functions to get responsible output. We validate our proposed approach on both *ex-vivo* and *in-vivo* laparoscopic images where we are able to reach improved results over existing state-of-the-art approaches on the same dataset.

The remaining sections of the literature are organized as follows: Section 2 is dedicated to the previous state-of-the-art related works. Section 3 is for the description of the methodologies and datasets. In section 4, the different experiments and results are reported. All the results are interpreted in section 5. Finally, thesis is concluded in section 6.

2. State of the art

The previous state-of-the-art works related to this thesis work have been described in three sub-sections as follows:

Surgical Instrument Detection: The pioneer vision-based methods used color-marker segmentation via low-level image processing to detect the shaft or the tip of the instrument (Krupa et al., 2003; Wei et al., 1997). Those methods are accurate in tracking and efficient in computation but suffer from unresolved issues of color and lighting variations. Some other vision-based methods exploit the geometric constraints and the gradient-like features, in order to identify the shaft of instrument, but fail to provide more accurate 3D positions of the instrument tip (Agustinos and Voros, 2015). As well as, in some images where the edge of the instrument almost blended with the background tissue, the gradient-like features often fail to detect the instruments. For detecting any kind of object, CNN already achieved the benchmark over other computer vision algorithms (Arel et al., 2010). Jin et al. (2018) developed an approach by leveraging region-based convolutional neural networks (R-CNNs) to perform spatial detection of surgical instruments. But, they were able to obtain 5 fps as a processing speed which is very less for real-time AR based MIS applications. Choi et al. (2017) applied unified architecture YOLO to detect the surgical instrument and their reported precision was only 72.26%.

Surgical Instrument Segmentation: Allan et al. (2013) proposed probabilistic supervised classification method to detect pixels belonging to surgical instruments using RF and used variable importance to pick four color features such as hue, saturation from HSV color space and opponent 1 and opponent 2. Finally, RF was trained on those features to classify the pixels in the test set with no post-processing stages. Agustinos and Voros (2015) used color and shape information of the surgical instrument for the segmentation which was based on the CIELab color space, a grayscale image composed of the a and b channels, corresponding to the chromaticity, Cab. After that, post-processing consisting in automatic Otsu thresholding and a skeletonization was used and followed by an morphological erosion. Finally, a contour detection algorithm (Suzuki and Be, 1985) was used to extract the extreme outer surgical instrument's contour of each region as an oriented bounding box. Garcia-Peraza-Herrera et al. (2017b) proposed a real-time automatic method based on Fully Convolutional Networks (FCN) with the improved learning process. To improve the learning process, they used Cyclical Learning Rate (CLR) (Smith, 2015) where the LR boundaries, momentum, and weight decay were $[1e-13, 1e-10]$, 0.99 and 0.0005 respectively. For better optimization of the cost function, they employed the pre-trained model which was trained on PASCAL-context 59-class (60 including background) (Mottaghi et al., 2014) dataset. Attia et al. (2017) applied a hybrid CNN method utilizing both recurrent and convolutional networks simultaneously. To prevent the coarse segmentation (Pakhomov et al., 2017), Recurrent Neural Network (RNN) was trained to model contextual relationships between pixels. In their network, four layers of the recurrent neural network are employed to find local/global dependencies between pixels in coupled directions. Garcia-Peraza-Herrera et al. (2017a) proposed two novel deep learning architectures for automatic segmentation of non-rigid surgical instruments namely ToolNetMS and ToolNetH. In ToolNetMS, all scales in FCN8s were summed in a cascaded fashion to ensure that the responses around the edges are better than traditional FCN8s. On the other hand, in ToolNetH, instead of element-wise sum, aggregation of all *sigmoid* cross-entropy losses from multiple scales has been performed after inspiration from holistically-nested edge detection (Xie and Tu, 2015). Pakhomov et al. (2017) employed deep residual learning and dilated convolutions for the instrument segmentation. To mitigate the coarse output from the segmentation network, they used strides equal to one in the last two convolutional layers which are responsible for downsampling in ResNet-101 (He et al., 2016) and by employing dilated convolutions for subsequent convolutional layers. Their proposed architecture was originally designed for edge detection problems which are almost similar to FCN8s (Shelhamer et al., 2017).

Pose Estimation of Surgical Instrument: Most of the image-based methods for the surgical pose estimation extracted low-level visual features from key-points or Regions of Interest (ROI) to learn offline or online part appearance templates by using machine learning algorithms (Rieke et al., 2016). Such a low-level feature of the surgical instrument usually suffers from a lack of semantic interpretation of the pixel and cannot capture the high-level category appearance (Du et al., 2018). Laina et al. (2017) reformulated the 2D instrument pose estimation as heatmap regression and was able to get concurrent, robust, and near real-time regression of localization and segmentation of the surgical instrument via deep learning. Li et al. (2014) proposed Instrument Tracker via Online Learning (ITOL) for the instrument track where ITOL used a robust gradient-based tracker which was capable of failure detection as the basic tracker. But, it was not detecting forceps tips and gradient-based approaches often fail to track due to having very less edge information in some images. Reiter et al. (2012) proposed tracking approach of the surgical instrument using the landmarks on its body surface where color, location, and gradient-based features have been associated with the landmarks. In their method, localization had been done by matching the features tracks in the stereo camera using normalized cross-correlation with a high degree of localization accuracy. But, the computational cost of extracting all of those features is huge and the occlusions of some landmarks due to the instrument rotation might result in a high degree of localization error (Alsheakhali et al., 2016). Allan et al. (2015) proposed articulated instrument tracking in 3D laparoscopic images where color information was used for the multi-class segmentation. The 3D pose of the instrument estimated from different statistical models as well as the optical flow was used for pose tracking from image to another. Expensive feature extraction and high sensitivity to the light changes (Alsheakhali et al., 2016) are the drawbacks of their method. Rieke et al. (2015) proposed a regression forests model to localize the forceps tips within a bounding box but that bounding box was provided using intensity-based tracker which was not robust and there is the possibility of losing the tracker.

In this thesis, we propose ART-Net which has five sub-network branches for instruments detection, segmentation, and three geometric features i.e. edge-line (tool boundary), mid-line, and tool-tip (see Table 1) extraction for the $3D$ pose estimation using algebraic geometry. All the sub-networks share the same convolution network which has five blocks and thirteen layers that mimic the VGG-16 (Simonyan and Zisserman, 2014). To make robust and lightweight SIMO network for real-time AR applications in MIS, we use depth-wise separable convolution in the decoder of segmentation and regression sub-networks which is in-

spired by Xception network (Chollet, 2017) and work of Kaiser et al. (2017). To retrieve the lost spacial and edge information due to sub-sampling, to reduce checkerboard noise and to minimize over-segmentation, we have added one special skip connection with traditional skip connection in U-Net. The new skip connection concatenates at the very last layer of decoder through a bunch of depth-wise separable convolutions without sub-sampling of the original image to produce the full resolution feature map and termed as Full resolution Feature map Generator (FrG). For the segmentation sub-network, we have proposed to combine $cross\ entropy$ and IoU as a cost function to obtain maximum overlapping between predicted and true surgical instrument mask.

3. Material and methods

In this section, the proposed pipeline, the architectural design of ART-Net, steps for 3D pose estimation using algebraic geometry, and used datasets are presented. The pictorial presentation of the overall pipeline is shown in Fig. 2. Firstly, input images are fed to CRB (Conv+ReLU+Batch Normalization (BN)) block which is a bunch of convolutions, *relu* activation, and batch normalization with sub-sampling and is shared by all other sub-networks for detection, segmentation, and regression simultaneously. FCS (Fully Connected+Softmax) is the fully connected layer and followed by the $softmax$ classifier that provides the information about the tool present in the image frame and termed as detection sub-network. If the detected tool flag is yes from the detection sub-network, the pose of the surgical tool will be estimated and vice-versa. DRBS (Deconv+ReLU+BN+Sigmoid) named as segmentation sub-network and DRB (Deconv+ReLU+BN) named as regression sub-network is the bunch deconvolution for semantically tissue or instrument pixel labeling and the regression to get three geometric features for pose estimation respectively. The methodologies of designing the proposed network, 3D pose estimation, and annotation are elaborately described in the following sub-sections.

3.1. Detection, Segmentation, and Geometric Feature Extraction sub-network

In general, CNN for the semantic segmentation consists of two essential parts for the pixel-wise classification (Badrinarayanan et al., 2017). The first part is encoder which is composed of convolution layers and sub-sampling layers where convolution layers are responsible for the automatic features extraction (Lin et al., 2013). The purpose of the sub-sampling layers is to achieve spatial in-variance by reducing the resolution of the feature maps and also to improve the robustness of the classifier due to the elimination of the redundant features. Sub-sampling also increases the field of view of the feature maps to extract more abstract class salient features and minimizes computation time (Shelhamer et al., 2017). On the other hand, the second part is a decoder which semantically projects the discriminating features of lower resolution learned by the encoder onto the pixel space of higher resolution to get a dense pixel-wise classification (Garcia-Garcia et al., 2018). In semantic segmentation, the encoder part is quite similar to all the CNN models but they mainly differ in the decoder mechanism. Semantic segmentation not only requires discrimination at pixel level but also a decoder mechanism to project the discriminating features learned at different stages of the encoder onto the pixel space. However, the significantly reduced feature map due to sub-sampling suffers from spatial resolution loss which introduces coarseness, less edge information, checkerboard noise, and over-segmentation in semantically segmented mask (Odena et al., 2016; Ronneberger et al., 2015; Shelhamer et al., 2017). When the kernel size of the deconvolution is not divisible by the up-scaling factor, the number of low-resolution features that contribute to a single high-resolution feature is not constant across the high-resolution feature maps which is called *deconvolution overlap* and is one of the causes of checkerboard artifact in the segmented mask.

To overcome those limitations, Ronneberger et al. (2015) introduced skip connection in their popular U-Net which allows the decoder at each stage to learn back relevant features that are lost when pooled in the encoder. Shelhamer et al. (2017) fused features from different coarseness to refine the segmentation using spatial information from different resolutions at different stages from the encoder. But, in that model, there is a dependency between kernel size and up-sampling factor to avoid *deconvolution overlap*. Al-masni et al. (2018) proposed Full resolution Convolution Network (FrCN) which does not have any sub-sampling in the encoder to preserve the spatial information of the feature map for precise segmentation. But, sub-sampling of feature map has several positive aspects in CNN as mentioned at the beginning of this sub-section. Pakhomov et al. (2017) employed dilated (atrous) convolutions to enable initialization with the parameters of the original classification network. However, to overcome the sub-sampling limitations and deconvolution overlap, we have employed two types of skip connections. The first one is between the corresponding same dimensional feature map in both encoder and decoder which has ladder-like structure (Rasmus et al., 2015) and is inspired from U-Net. The second one so-called FrG that connects the very end layer of the decoder with the original image via a stack of depth-wise separable convolution without sub-sampling to provide the full resolution feature map which is a compensatory to the lost spacial information

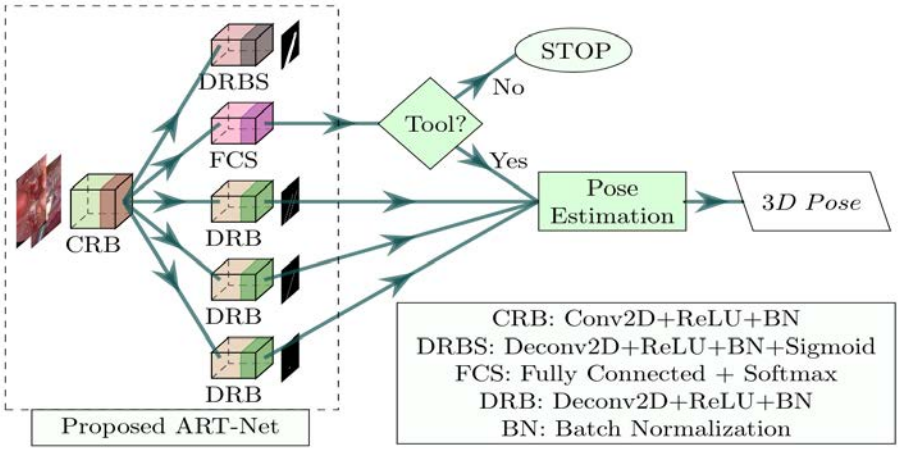


Figure 2: The pipeline of the proposed framework for concurrent detection, segmentation, and geometric feature extraction for 3D pose estimation of the surgical instrument. The features from three regression blocks (DRB) and the tool flag (yes or no) from FCS block are used for the pose estimation and the segmented mask is used for occlusive visualization of the surgical tool during the occlusion by other organs.

due to sub-sampling.

In traditional CNN based image classification networks, *flatten* layer is used to vectorize the $2D$ arrays into a single long continuous linear vector where pooled feature map flattened into a sequential column of numbers and followed by several densely connected layers. So, most of the parameters of such classifiers belong to the fully connected layers and can cause the overfitting of the classifier. To overcome that limitation, a *Dropout* (Srivastava et al., 2014) layer is used as a regularizer which randomly set half of the activation to the fully connected layers to zero during training. Thus, it improves the generalization ability and largely prevents overfitting of the CNN based classifiers. Lin et al. (2013) proposed a global average pooling (GAP) layers where only one feature map is generated for each corresponding category. GAP sums out the spatial information of the incoming pooling layer, thus it provides more robustness to spatial translations of the input. GAP layers also perform a more extreme type of dimensionality reduction to avoid overfitting. In GAP, $height \times weight \times depth$ dimensional tensor reduced to $1 \times 1 \times depth$ where each $height \times width$ feature map transfer to a single number by simply taking the average of all $height.width$ values. In our detection subnetwork, we used GAP instead of the traditional $flatten$ layer due to having the state-of-the-art performance for image classification. We also used $dropout$ followed by $softmax$ classifier to detect the surgical tool in the incoming laparoscopic frames. The use of GAP also provides lightweight detection sub-network which is beneficiary for the SIMO ART-Net. For the simplicity of presentation, we have divided the whole ART-Net structure into two parts as shown in Fig. 3 but the complete structure of ART-Net is available in GitHub[1].

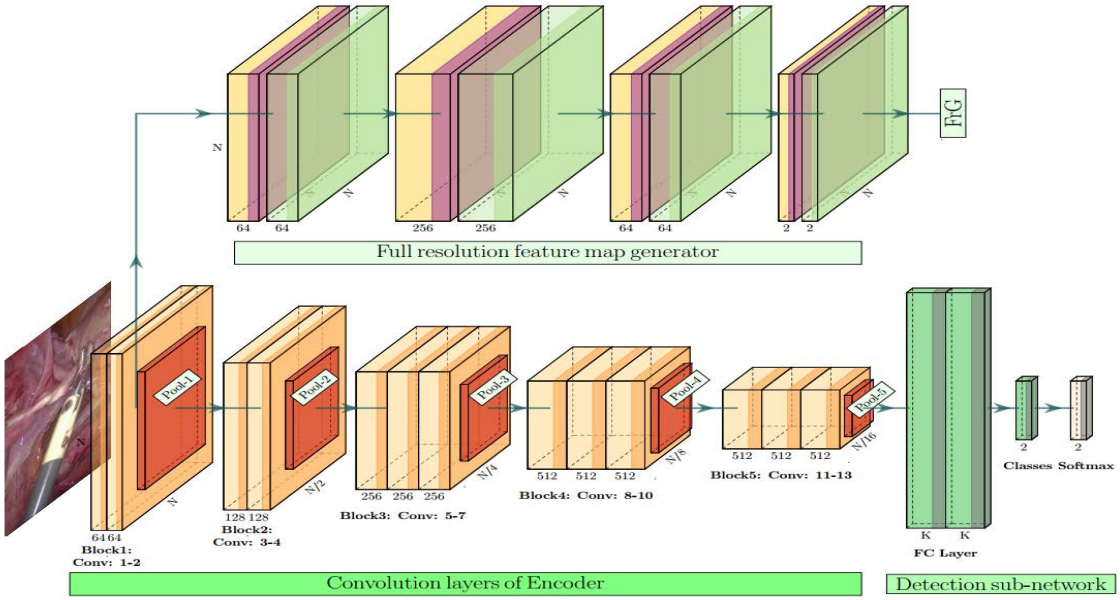
The lightweight ART-Net is achieved by using depth-wise separable convolution (Chollet, 2017) instead of using traditional standard convolution. Depth-wise separable convolution is a spatial convolution performed independently over each channel of input and followed by a point-wise convolution i.e. a 1×1 convolution where projecting the output of the channel by the depth-wise convolution onto a new channel space. For any convolution layer, if we have F numbers of filter, M depths, and D_K kernel size, the total numbers of parameters will be $F \times M \times D_K^2$ and $M \times (F + D_K^2)$ for standard and depth-wise separable convolution respectively. Thus, we are able to reduce the number of parameters by $(1/F + 1/D_K^2)$ times of any convolution layer of proposed ART-Net.

3.2. 3D Pose Estimation

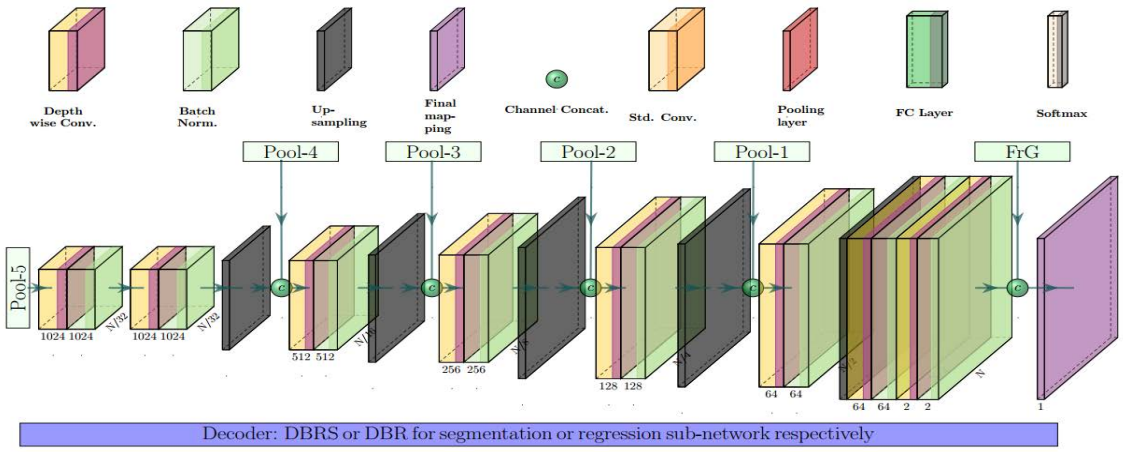
The rotation $R \in SO(3)$ which denotes the direction and translation $t \in \mathbb{R}^3$ which relates the camera position of the instrument pose, $[R \mid t]$ is estimated by geometric solver using sets of geometric features obtained from ART-Net. The rotation, r_3 of $R = [r_1 \; r_2 \; r_3]$ is directed to the cylindrical axis as shown in Fig. 4 and r_1 is the normal to the plane APB. r_2 is perpendicular to both r_1 and r_3 and is directed to the camera optical axis (OA). The proposed pipeline for 3D pose estimation is presented below step-by-steps.

Step 1: The constrain functions for the homogeneous measurement matrix are ${L_1}^T V = 0$, ${L_2}^T V = 0$, and ${L_3}^T V = 0$ where L_1 and L_2 are two lines of edge line, L_3 is the mid-line, and V is the vanishing point and the homogeneous measurement matrix can be expressed as below.

$$\begin{bmatrix} L_1 \\ L_2 \\ L_3 \end{bmatrix}^T V = 0 \implies A^T V = 0 \qquad (1)$$



(a) Part-1: Encoder, Full resolution Feature map Generator (FrG) and Detection sub-network of ART-Net



(b) Part-2: Decoder for segmentation or regression sub-networks of ART-Net

Figure 3: Different parts of the proposed ART-Net. Part-2 is replicated four times for segmentation, and three regression (three geometric features) sub-networks to construct complete ART-Net.

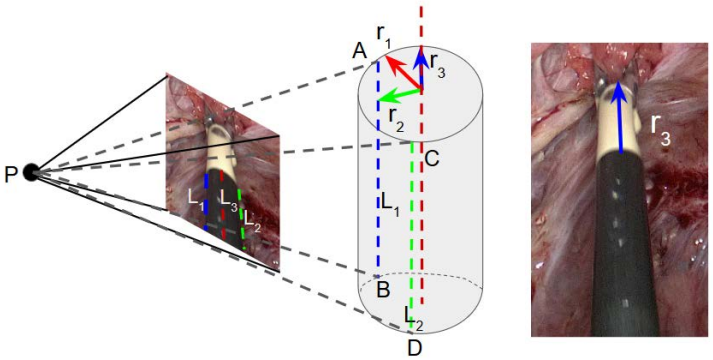


Figure 4: Pictorial presentation of perspective projection of the surgical instrument and illustration of 3D pose estimation of the surgical tool where tool is considered as the cylindrical object with radius r. P is the optical camera center which forms planes APB and CPD with the edge line L_1 and L_2 respectively. To show the direction of the rotation R, RGB color convention has been used.

The vanishing point, V can be calculated using the singular value decomposition (SVD) from the homogeneous measurement matrix, A.

Step 2: The rotation around the $z-axis$ (instrument axis) is calculated from the vanishing point, V and camera intrinsic parameter matrix, K as follows.

$$r_3 = \frac{K^{-1}V^T}{\left\|K^{-1}V^T\right\|} \quad where,\ K = \begin{bmatrix} f & 0 & c_x \\ 0 & f & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

where f is the focal length and (c_x, c_y) is the optical center of the endoscopic camera.

The rotation around the $x-axis$, r_1 is the solutions of outer product, $r_3 r_3^T$ of r_3 and the rotation around the $y-axis$, r_2 is the cross product of r_1 and r_3, $r_2 = r_3 \times r_1$.

Step 3: The translation along the cylinder axis where the translation is the locus of points belonging to both planes $P1$ and $P2$. $P1$ is the normalized median plane between the planes through the L_1 and L_2 with optical center and $P2$ is normalized plane through the cylinder axis whose normal is directed towards the optical center.

Step 4: Finally, we refine the pose by nonlinear optimization of the residual between initial tool boundary, mid-line, and tip-point with predicted geometric features from ART-Net.

3.3. Datasets

To demonstrate the performance of the proposed network, two different datasets have been used. The first one is MICCAI 2015 Endoscopic Vision Challenge - Instrument Segmentation and Tracking Subchallenge (MICCAI, 2015) which contains two subdatasets namely: robotic (non-rigid) and non-robotic (rigid). The robotic and non-robotic data has the resolution of 720×576 and 640×480 respectively. The second one is our annotated data (see GitHub[1]) where for each image, we annotated a binary mask, three geometric features namely edge-line (tool boundary), midline, and tip-point as shown in Table 1. The summary and distribution of the two different datasets used in this thesis are provided in Table 2.

3.4. Training and Run-time Analysis

The kernels in encoder of the proposed ART-Net were initialized with the pre-trained weights of VGG-16 (Simonyan and Zisserman, 2014) trained on ImageNet (Deng et al., 2009) whereas the kernels in decoder part were initialized with the *glorot_uniform* distribution. We have performed two stages of training and testing where in the *stage-1*, we trained and tested only the segmentation sub-network on EndoVis (robotic) data and in *stage-2*, we trained and tested whole ART-Net on EndoVis (non-robotic) + our annotated data. *Stage-1* is dedicated for the evaluation of proposed segmentation sub-network of ART-Net for the surgical instrument segmentation and compare against recent state-of-the-art tool segmentation networks on the same dataset (EndoVis-robotic). In *stage-1*, we also trained and tested standard U-Net and FCN8s on the same dataset (EndoVis-robotic) to compare against our proposed segmentation sub-network. *Stage-2* is employed for getting the concurrent output from the ART-Net where the EndoVis (non-robotic) dataset was further annotated as like as Table 1 and merged with our annotated data since both are non-robotic (rigid) data.

The segmented mask from segmentation sub-network has been evaluated using mean Dice-similarity coefficient (mDSC), mean Sensitivity (mSn.), mean Specificity (mSp.), mean Balanced Accuracy (mBA), and mean Intersection over Union (mIoU). mDSC and mIoU are the indicator for the amount of overlapping between the predicted and true mask whereas mSn., mSp., mBA are the metrics for evaluating the true positive rate and false positive rate of the predicted mask. The predicted edge and mid-line features are quantitatively evaluated using the mean Arc Length (mAL) error of the unit circle between true point x_{GT} and predicted point $\hat{x}_P$ as shown in Fig. 5 and can be expressed as Eq. 3. The predicted instrument tip-point is evaluated using euclidean distance between true point and predicted point.

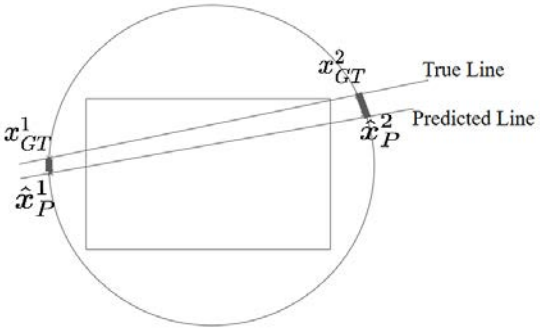


Figure 5: Arc Length (AL) error measurements for the evaluation of predicted edge and mid line features from regression sub-network of ART-Net. x^1_{GT} and x^2_{GT} are two cross points of the true line whereas $\hat{x}^1_{GT}$ and $\hat{x}^2_{GT}$ are two cross points of the predicted line.

$$mAL = \frac{1}{N} \times \sum_{i=1}^{N} \frac{{d_i}^1(x^1_{GT}, \hat{x}^1_P) + {d_i}^2(x^2_{GT}, \hat{x}^2_P)}{2} \tag{3}$$
$$where,\ d\ is\ arc\ length\ of\ unit\ circle$$

where N is the total number of images. For both classification and semantic segmentation, widely used and stable loss function is binary or categorical cross-

Table 1: Example of our annotated image along with the corresponding binary mask, edge-line, mid-line, and tip-point. Annotation has been achieved using the ImageJ software and basic image processing methods (see GitHub[1]).

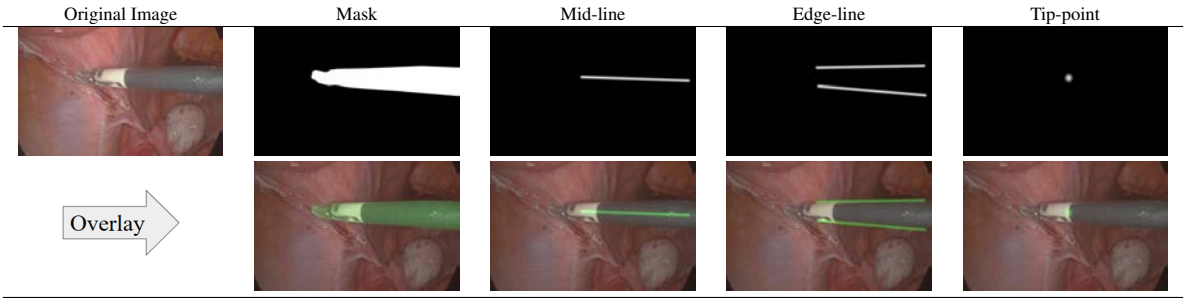


Table 2: Summary and distribution of the EndoVis-2015 (both robotic/non-rigid and non-robotic/rigid) and our annotated data (see GitHub[1]). To get the frames from EndoVis (robotic) video, FFmpeg cross-platform was used.

Train Data	Robotic	**Dataset-1**	**Dataset-2**	**Dataset-3**	**Dataset-4**	-	-
		$25fps \times 45s$	$25fps \times 45s$	$25fps \times 45s$	$25fps \times 45s$	-	-
	Non-robotic	**OP1**	**OP2**	**OP3**	**OP4**	-	-
		40	40	40	40	-	-
Test Data	Robotic	**Dataset-1**	**Dataset-2**	**Dataset-3**	**Dataset-4**	**Dataset-5**	**Dataset-6**
		$25fps \times 15s$	$5fps \times 15s$	$25fps \times 16s$	$25fps \times 15s$	$25fps \times 61s$	$25fps \times 60s$
	Non-robotic	**OP1**	**OP2**	**OP3**	**OP4**	**OP5**	**OP6**
		10	10	10	10	50	50
Our Data	Train Data	**Images**	**Masks**	**Edge-lines**	**Mid-lines**	**Tip-points**	-
		508	508	508	508	508	-
	Test data	**Images**	**Masks**	**Edge-lines**	**Mid-lines**	**Tip-points**	-
		127	127	127	127	127	-

entropy. But, for the surgical tool segmentation, the background tissue differs from the surgical tool which may have a biased effect on one particular class (Garcia-Peraza-Herrera et al., 2017a). The proposed cost function, L_{seg} for the segmentation is the sum of binary cross-entropy and intersection over union (Yuan and Lo, 2019) which can be expressed as Eqn. 4.

$$L_{seg}(y,\hat{y}) = \frac{1}{N}\sum_{i=1}^{N}[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)] + 1 - \frac{\sum_{i=1}^{N} y_i \times \hat{y}_i}{\sum_{i=1}^{N} y_i + \sum_{i=1}^{N} \hat{y}_i - \sum_{i=1}^{N} y_i \times \hat{y}_i} \quad (4)$$

where y and $\hat{y}$ are the true label map and predicted probability map respectively. The cost function for the detection and regression sub-network is *cross entropy* and *mean squared error* respectively. The total cost function, L_{SIMO} of ART-Net is the weighted sum of the individual cost function of each sub-network and is calculated using Eqn. 5.

$$L_{SIMO}(y,\hat{y}) = W_{seg} \times L_{seg}(y,\hat{y}) + W_{det} \times L_{det}(y,\hat{y}) + W_{mid} \times L_{mid}(y,\hat{y}) + W_{edge} \times L_{edge}(y,\hat{y}) + W_{tip} \times L_{tip}(y,\hat{y}) \quad (5)$$

where W is the scalar weight of the cost function corresponding to individual sub-network and is the portion of L_{SIMO}. The cost function, L_{SIMO} is optimized using *adadelta* (Zeiler, 2012) with *initial learning rate* = 1.0 and *decay factor* = 0.95.

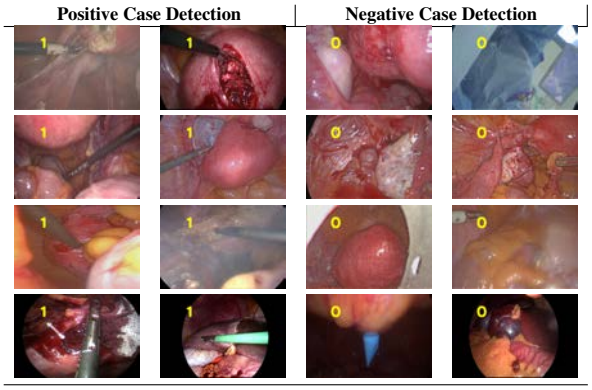
4. Experiments and Results

The experimental results for detection, segmentation, geometric feature extraction, 3D pose estimation, and applications to augmented reality are reported in several sub-sections as follows:

4.1. Instrument Detection

The instrument detection sub-network of ART-Net provides information about the instrument presence in the laparoscopic image frame and generates instrument's FLAG. No instrument presence in the image frame provides FLAG=0 which is named as negative case whereas instrument presence in the image frame provides FLAG=1 which is named as positive case. The performance of the detection sub-network of ART-Net has been evaluated using average precision, average accuracy, and area under the ROC (receiver operating characteristic) curve (AUC). The obtained average precision, average accuracy, and AUC are 100.0%, 100.0%, and 1.0 respectively for the detection. The reported quantitative metrics for the surgical instrument detection demonstrates an impressive performance of proposed surgical instrument detection sub-network over the state-of-the-art. A few qualitative results for the detected positive and negative cases of the surgical instrument are shown in Table 3.

Table 3: The qualitative results for the surgical instrument detection from ART-Net. Left two columns are for the positive case detection where instrument's FLAG=1 (instrument present) whereas the right two columns are for negative case detection where instrument's FLAG=0 (instrument not present). The more qualitative results for instrument detection are available in GitHub[1].



From the qualitative results of the instrument detection, it is worth to mention that our detection sub-network of ART-Net can distinguish the texture information of the background tissue pixels and instrument pixels although there is more complex texture information in background tissue than surgical instrument. It is also seen that the detection sub-network can detect the surgical instrument even in the presence of different types of noise and instrument edges are almost blended with background pixels. The trocar (4^{th} row - 3^{rd} column) and operation theater (1^{st} row - 4^{th} column) are also detected as the negative case accurately.

4.2. Instrument Segmentation

The segmentation obtained from the proposed ART-Net, U-Net, and FCN8s has been evaluated both quantitatively and qualitatively. The quantitative and qualitative results of the instrument segmentation are shown in Table 4 and Table 5 respectively. The results provided for both quantitative and qualitative judgment are without any post-processing.

From the qualitative results of the surgical instrument segmentation as shown in Table 5, it is seen that the segmented mask generated from FCN8s has poor edge information (instrument boundary), checkerboard noise, and more false positive rate (FPR). But for 3D pose estimation from the vanishing point using geometric features, the edge information needs to be as precise as possible. The U-Net model has better quantitative segmentation results and has better edge response than FCN8s. But, in the noisy laparoscopic images where spatial information of the tool is not well visible, our proposed ART-Net performs better than any other networks by adding more spatial information at the very end of the decoder. The state-of-the-art mDSC and mIoU of the segmented mask from ART-Net have demonstrated a high degree of overlap between the true and predicted instrument mask. The quantitative and qualitative results for instrument segmentation in *Stage-1* as shown in Table 4 and Table 5 respectively have proven that our proposed segmentation sub-network provides more better segmented mask even in noisy laparoscopic images than U-Net, FCN8s, and the state-of-the-art networks. So, for the building of complete ART-Net, we will consider our proposed segmentation sub-network for concurrent regression and segmentation in *stage-2*. The qualitative segmentation results from *stage-2* on combined EndoVis (non-robotic) and our annotated data are shown in Table 6. The mDSC, mSn., mSp., mBA, and mIoU of the segmentation in *stage-2* are 93.22%, 95.30%, 98.98%, 97.14%, and 88.22% respectively. Those quantitative metrics have proven that there is very less true negative (instrument pixels as background tissue pixels) and false positive (background tissue pixels as instrument pixels) on the segmented mask of the surgical instruments in *Stage-2*. From Table 6, is seen that the segmentation sub-network provides precise segmentation of the surgical instrument with sharper edge response although the images are noisier than EndoVis (robotic). Both the quantitative and qualitative results for the segmentation using segmentation sub-network of ART-Net have demonstrated its outstanding performances comparing the state-of-the-art.

4.3. Geometric Features Extraction

The predicted geometric features (edge-line, mid-line, and tip-point) from the regression sub-networks of ART-Net have been evaluated quantitatively using the mean and median value of AL (see sub-section 3.4) in degree and euclidean distance in pixels. The qualitative results for the predicted geometric features are shown in Table 7. The probability map of the geometric line features (edge and mid-line) are approximated using hough line detection by exploiting the duality between points on a curve and parameters of that curve (Ballard, 1981). The probability map of the tip-point is approximated by using simple $argmax2D$.

The mean and median AL for the edge-line and mid-line are $(2.45°, 1.71°)$ and $(2.23°, 1.34°)$ respectively whereas the mean and median euclidean distance for the tip-point is $(9.3, 3.2)$ pixels. The mean AL for edge-line and mid-line indicate that the angular displacement of the predicted line from the true line are very less compared to highest possible angular displacement. The median AL has denoted that 50% AL lies less than 1.71° and 1.34° for edge-line and mid-line respectively which has proved better robustness of the line feature detection by ART-Net. The mean, 9.3 and median, 3.2 value of euclidean distance in pixels between true and predicted tip-point shows the success of tip-point regression sub-network of proposed ART-Net. From Table 7, it is seen that the predicted (Green color) edge-line and mid-line are almost overlapping the true lines (Yellow color). Both the qualitative and quantitative results of

Table 4: Segmentation performance metrics from ART-Net, U-Net, FCN8s, and recent state-of-the-art networks on EndoVis-2015 (robotic) for quantitative assessment and comparison. Metrics for our proposed and implemented networks were calculated using the true labels and semantic labels obtained from network.

Networks	Pre-train	Datasets	Performance Metrics				
			mDSC	mSn.	mSp.	mBA	mIoU
FCN (Garcia-Peraza-Herrera et al., 2017b)	PASCAL-context	EndoVis (Robotic)	-	72.2%	95.2%	83.7%	-
ToolNet (Garcia-Peraza-Herrera et al., 2017a)	NA	EndoVis (Robotic)	82.2%	-	-	81.0%	74.4%
FCN (Pakhomov et al., 2017)	PASCAL VOC	EndoVis (Robotic)	-	85.7%	**98.8%**	92.3%	77.6%
U-Net	ImageNet	EndoVis (Robotic)	87.5%	**93.5%**	97.5%	93.3%	78.1%
FCN8s	ImageNet	EndoVis (Robotic)	86.4%	85.9%	98.3%	92.1%	76.5%
ART-Net	ImageNet	EndoVis (Robotic)	**89.3%**	88.1%	98.6%	**93.4%**	**81.0%**

Table 5: Qualitative segmentation results of ART-Net, U-Net, and FCN8s on EndoVis-2015 (robotic). Same test images were shown for all networks to compare qualitative performance. The DSC has been also added with the overlaid image to verify the segmentation accuracy quantitatively. The more qualitative results of the segmented mask in *Stage-1* are available in GitHub[1].

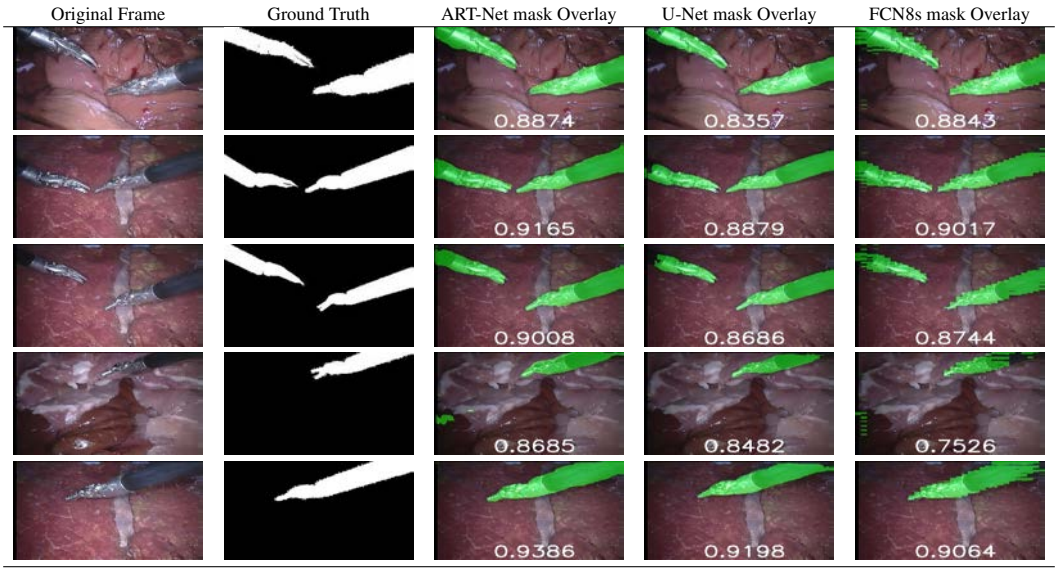
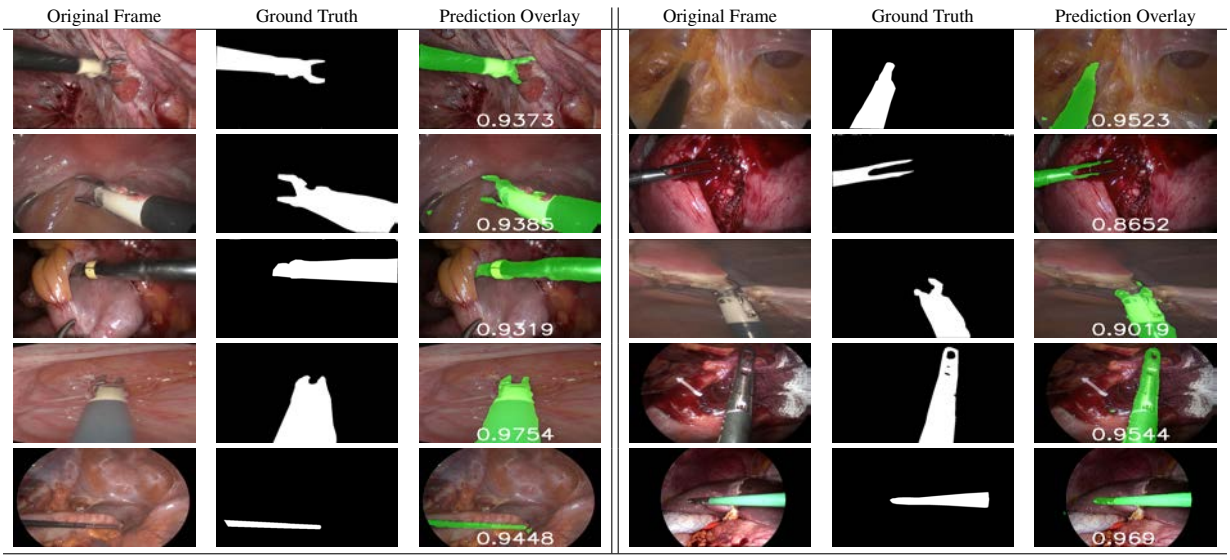
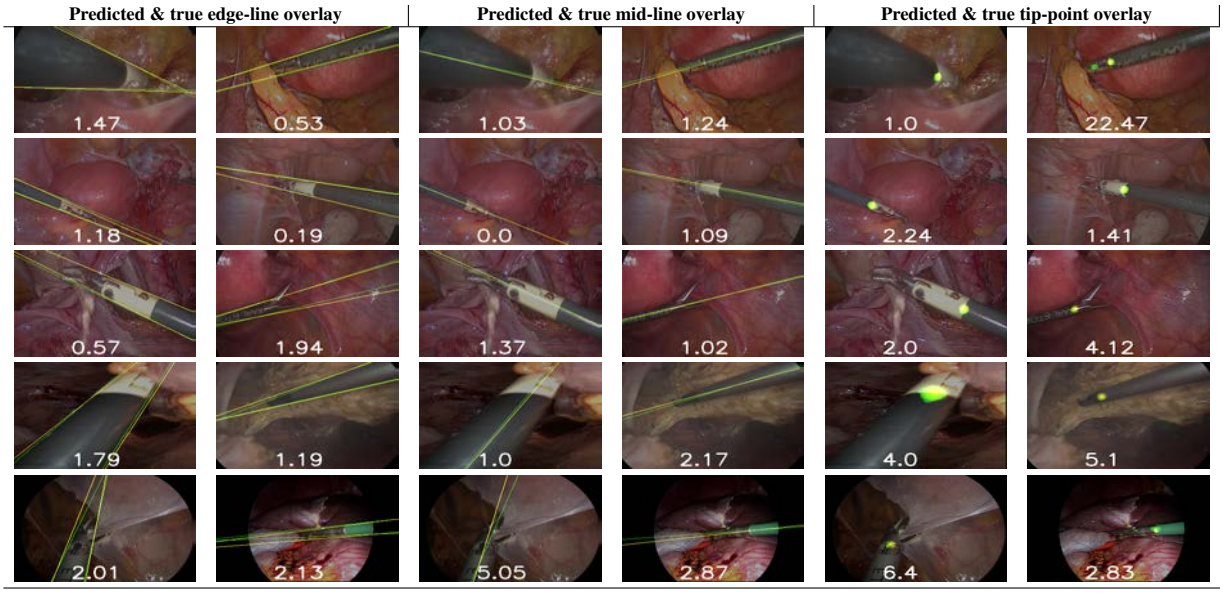


Table 6: Qualitative segmentation results from *Stage-2* training and testing on EndoVis (non-robotic) + our annotated data. The DSC has been also added with the overlaid image to verify the segmentation accuracy quantitatively. The more qualitative results of the segmented mask in *Stage-2* are available in GitHub[1].



the geometric feature extraction have validated the excellent performance of regression sub-networks of the

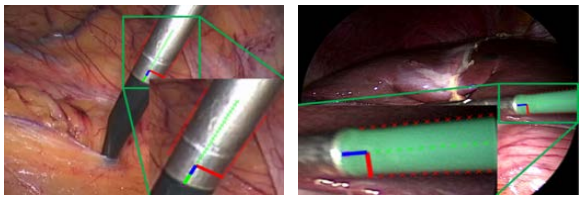
Table 7: Approximated geometric features from the ART-Net prediction (Green color) and true features (Yellow color) are overlaid with the original image frame. The arc length measured in *degree* and euclidean distance measured in pixels are also added with the overlaid images for quantitative assessment. The more qualitative results for the approximated geometric features are available in GitHub[1].



proposed ART-Net.

4.4. 3D Pose Estimation

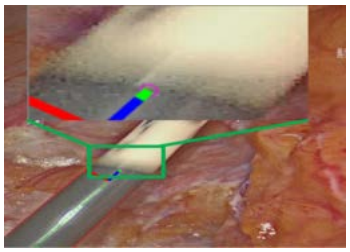
The 3D pose of the surgical instrument has been estimated according to the steps described in 3.2 from the sets of geometric features obtained from proposed ART-Net. The estimated 3D pose of the surgical instruments from proposed approach are shown in Fig. 6 for qualitative assessment, where it is seen that the estimated instrument pose is irrespective of instrument colors and the navigation directions of the instrument. From Fig. 6, it is also seen that the estimated 3D pose of the instrument is accurate even the instrument diameter and head length are different. The more qualitative accuracy of our approach for 3D pose estimation is shown in Fig. 7, where it is seen that estimated pose is precise even there is a motion blur between the surgical instrument body and instrument head (see Fig. 7 (a)). From the gradient magnitude using 1st derivative (see in Fig. 7 (b)) and 2nd derivative (see in Fig. 7 (c)), it is seen that there is no gradient information at the boundary of the instrument body and instrument head. So, using the traditional local filtering, it is often impossible to obtain gradient information about the tool head whereas our approach is successfully locating the tip-point using the features from proposed ART-Net. In Fig. 7 (d) and Fig. 7 (e), it is demonstrated that there is no gradient information along the tool edge (tool boundary) using the local filtering (sobel), but our CNN based method can locate the instrument boundary and estimate the 3D pose of the instrument precisely. It is also demonstrated in Fig. 7 (d) and Fig. 7 (e) that after non-linear refinement, the contour of the surgical tool is exactly at the maximum gradient location which is often impossible to achieve using any other local filtering of the traditional image analysis. From Fig. 7 (g) and Fig. 7 (h), it is also noticeable that there is no edge or instrument tip information after applying local filtering due to motion blur of the instrument. But, our CNN based feature extraction and geometric solver is able to estimate the 3D pose (see Fig. 7 (f)) accurately. From the above discussions and figures, it is worth to mention that our approach for 3D pose estimation of the surgical instrument is robust and accurate even in noisy, blurred, and motion blurred laparoscopic images.



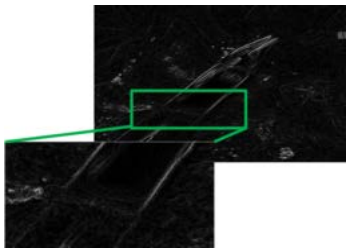
(a) Top to bottom directed tool (b) Right to left directed tool

(c) Left to right directed tool (d) Bottom to top directed tool

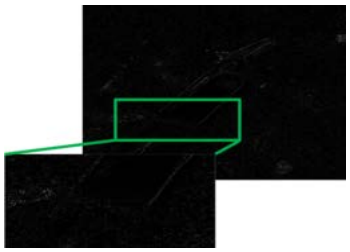
Figure 6: Representative examples of estimated 3D surgical instrument pose. For better visualization of the pose, selected ROI is zoomed. Red, green, and blue color denote the direction of r_1, r_2 and r_3 of the instrument's orientation (rotation) respectively.



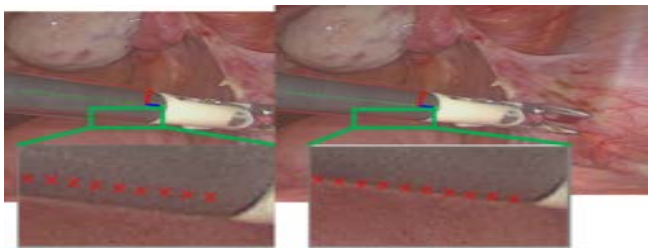
(a) Precise tip-point detection even there is motion blur



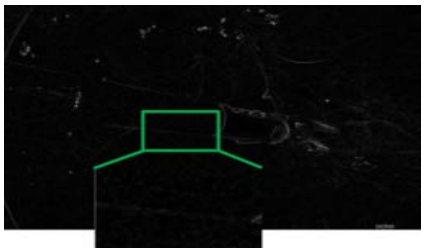
(b) Gradient magnitude using 1st order filtering (sobel)



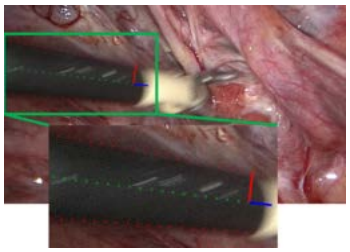
(c) Gradient magnitude using 2nd order filtering (laplacian)



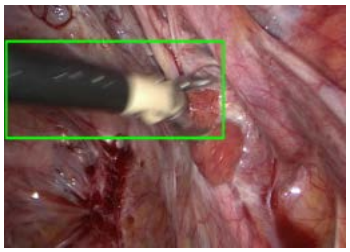
(d) Pose refinement for exact tool pose. (Left) before refinement and (right) after refinement



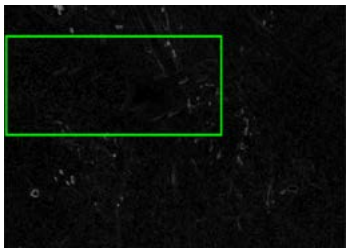
(e) Gradient magnitude using 1st order filtering (sobel)



(f) Precise 3D pose in complex case



(g) ROI selection using the segmented mask



(h) Gradient magnitude using 1st order filtering (sobel)

Figure 7: Representative examples of precise 3D pose estimation of the surgical instrument from our proposed pipeline using CNN based geometric features and geometric solver.

4.5. Applications to Augmented Reality

In this sub-section, we are presenting the practical implication of our segmented instrument mask and 3D pose of the instrument to AR in MIS. For occlussive visualization of the surgical instruments where the surgical instrument should not be occluded by the anatomy for better depth perception of the surgeon, we have used image blending in the region of the instrument mask (segmented mask). To avoid the aliasing effect during the blending, we used Eqn. 6 to blend the images smoothly.

$$I_{final} = (1 - \alpha) \times I_{aug} + \alpha \times I_{raw} \tag{6}$$

where, I_{final}, I_{aug}, and I_{raw} is the instrument occlusion augmented image, augmented image, and raw image respectively. The fraction coefficient, α is obtained from the predicted instrument mask from ART-Net after blurring with a 2D median filter. The occluded visualization of the augmented instrument is shown in Fig. 8 where the AR has been used with raw uterus image to locate the tumor inside uterus (see Fig. 8 (a)-(b) and Fig. 8 (d)-(e)). Fig. 8 (c) and Fig. 8 (f) of the occluded instrument where instrument occlude the anatomy have proved excellent performance of the segmentation sub-network of ART-Net. The proposed network also shows the worthy performance for both single and multiple instrument occlusion in AR on the uterus image. For more visualization, a supplementary video where the surgical instrument is not occluded by anatomy in AR on uterus images is available in YouTube[2].

5. Discussion

In this thesis, SIMO ART-Net has been proposed and implemented for surgical instrument detection, segmentation, and geometric feature extraction concurrently which was trained in an end-to-end fashion. From the results of the detection sub-network of ART-Net, it is observed that use of GAP in lieu of flattening layer has extreme dimension reduction capability which can

[2] `https://youtu.be/pAVYbapTbSc`

(a) Raw laparoscopic image (b) AR laparoscopic image (c) AR laparoscopic image with instrument occlusion

(d) Raw laparoscopic image (e) AR laparoscopic image (f) AR laparoscopic image with instrument occlusion

Figure 8: Application of the instrument segmented mask for instrument occlusion in augmented reality-based MIS on uterus images. Single instrument occlusion (a-c) and multiple instruments occlusion (d-f). A supplementary video of instrument occlusion in AR is available in the YouTube link[2].

provide more abstract features of the instruments to the CNN based image classifier. The dropout layer with GAP increases the generalization ability of the image classifier which has better robustness and can avoid the overfitting for surgical instrument detection. Additionally, the use of GAP also provides lightweight detection sub-network which can improve the detection rates and suitably applicable for real-time AR in MIS.

From the results of segmentation sub-network, it is seen that the use of skip connection has better potentiality of regaining the spatial information in the low-resolution image by learning back the relevant features that are lost when pooled in the encoder than fusing the features from the different coarseness of encoder. From the segmentation results of FCN8s (see Table 5), it is seen that there is a trade-off between the amount of up-sampling and the kernel size. The number of features is not constant at the border of the predicted instrument mask if the up-sampling scale and the kernel size are not divisible which introduces the checkerboard noise. As a consequence, the segmented instrument mask boundary is not smooth and there are more unwanted squares in the predicted mask (see Table 5). Compared to the FCN8s, U-Net provides better instrument mask due to skip connection. But, U-Net also fails to provide better instrument mask in noisy laparoscopic images due to having less spatial information as shown in Table 5. On the other hand, in our proposed ART-Net by introducing the FrG which concatenate more spatial information at the very end of the decoder, the segmented instrument masks are accurate and robust (see Table 4, Table 5, and Table 6). In the results of segmentation sub-network of ART-Net, it is also observed that the proposed cost function which is the summation of *cross entropy* and *IoU* performs better than any other cost functions for surgical instrument segmentation. Alone Cross entropy or IoU introduces more false positive and true negative respectively. Summation of *cross entropy* and *IoU* as a cost function reduces the false positive and true negative trade-off in segmented instrument mask.

The results of regression sub-networks for edge-line, mid-line, and tip-point of ART-Net demonstrated the outstanding performance of our proposed network both quantitatively and qualitatively. The L_2 loss function (Mean Squared Error) has excellent performance for regression than other regression loss functions (L_1 loss, quantile loss) due to having more stable and closed form solution. Furthermore, the use of FrG in regression sub-networks has great success for finding the instrument's features and sharper gradient at the instrument boundaries. Additionally, the non-linear *sigmoid* function at the end of the regression sub-networks truncates the probability between 0 to 1 and provides a probability map of the edge-line, mid-line, and tip-point.

From all the results for detection, segmentation, feature extraction of the surgical instruments, it is seen that the initialization of the kernels in the encoder with VGG16 weights trained on ImageNet is the better choice to avoid the falling in local minima during the

optimization of the cost function. From the outstanding results of ART-Net, it is worth to mention that as an optimizer *adadelta* also plays a crucial role which adapt learning rates based on a moving window of gradient updates instead of accumulating all past gradients.

The 3D pose of the surgical instrument was estimated using the geometric algebra and the geometric features from the regression sub-networks of ART-Net. To make the generic 3D pose estimation pipeline, only features are extracted from the ART-Net which are irrespective of instruments physical properties (instrument's diameter and head length). The estimated 3D pose from the vanishing point of the instrument boundary lines is robust as shown in Fig. 6 and Fig. 7. In some laparoscopic image where the instrument pixels are almost blended with the background tissue pixels near the instrument's edge, using the local computer vision filtering often impossible to get gradient information. But, CNN-based proposed approach can accurately find the gradient information at the instrument's edge which is irrespective of instrument's color, shape, size, orientation, etc. and can also estimate the instrument pose in case of motion blurred due to movement of instrument during the surgery.

6. Conclusions

In this thesis, the surgical instrument detection, segmentation, and 3D pose estimation using geometric features have been presented for an AR application in gesture guidance for laparosurgery. The proposed FrG has played a crucial role in compensating for the spatial information loss due to sub-sampling in segmentation and regression sub-networks of ART-Net. It can also be readily applicable to other kinds of encoder-decoder networks for semantic segmentation. In the case of SIMO like structure where the whole network comprises several sub-networks and trained in an end-to-end fashion, being lightweight is one of the core requirements for real-time applications. In the proposed ART-Net using depth-wise separable convolution and GAP, we were able to reduce the number of parameters approximately 3.6 times leading to a more generalized trained model all the while outperforming the state-of-the-art. Hence, the use of depth-wise separable convolution and GAP can be better choices for building lightweight SIMO type networks. Our approach for the geometric feature extraction is independent of the instruments physical properties and the estimated 3D pose is precise and robust. Hence, the proposed pipeline can be applied to any other clinical practice for 3D pose estimation of surgical instruments in the operation room (OR). The estimated 3D pose of the surgical tool is highly useful to solve the registration ambiguity without any extra tracking devices in the OR. The scale ambiguity of the structure from motion (SfM) can be solved by introducing this additional 3D pose information during the reconstruction of the pre-operative 3D model in AR based MIS. The occluded surgical instrument visualization in AR can overcome the depth perception limitation of the surgeon.

7. Acknowledgments

First of all, I would like to thank my supervisors Dr. Lilian Calvet and Prof. Dr. Adrien Bartoli for their crucial and inspiring suggestions during the thesis work. I am grateful to European Union and MAIA team for selecting me as an Erasmus funded student and also grateful to all of my class teachers during the whole MAIA studies. At last but not least, I am thanking to my family members, MAIA friends, and EnCoV members for their support in different aspects.

References

Agustinos, A., Voros, S., 2015. 2D/3D real-time tracking of surgical instruments based on endoscopic image processing. Computer-Assisted and Robotic Endoscopy 9515, 90–100. doi:10.1007/978-3-319-29965-5_9.

Al-masni, M.A., Al-antari, M.A., Choi, M., Han, S., Kim, T., 2018. Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks. Computer Methods and Programs in Biomedicine 162, 221–231. doi:10.1016/j.cmpb.2018.05.027.

Allan, M., Chang, P., Ourselin, S., Hawkes, D.J., Sridhar, A., Kelly, J., Stoyanov, D., 2015. Image based surgical instrument pose estimation with multi-class labelling and optical flow. International Conference on Medical Image Computing and Computer-Assisted Intervention 9349, 331–338. doi:10.1007/978-3-319-24553-9_41.

Allan, M., Ourselin, S., Thompson, S., Hawkes, D.J., Kelly, J., Stoyanov, D., 2013. Toward detection and localization of instruments in minimally invasive surgery. IEEE Transactions on Biomedical Engineering 60, 1050–1058. doi:10.1109/TBME.2012.2229278.

Alsheakhali, M., Eslami, A., Roodaki, H., Navab, N., 2016. CRF-based model for instrument detection and pose estimation in retinal microsurgery. Computational and Mathematical Methods in Medicine 2016, 592–600. doi:10.1155/2016/1067509.

Arel, I., Rose, D., Karnowski, T., 2010. Deep machine learning-a new frontier in artificial intelligence research. IEEE Computational Intelligence Magazine 5, 13–18. doi:10.1109/MCI.2010.938364.

Artal, R.M., Montiel, J.M.M., Tardos, J.D., 2015. ORB-SLAM: A versatile and accurate monocular slam system. IEEE Transactions on Robotics 31, 1147–1163. doi:10.1109/TRO.2015.2463671.

Attia, M., Hossny, M., Nahavandi, S., Asadi, H., 2017. Surgical tool segmentation using a hybrid deep cnn-rnn auto encoder-decoder, IEEE International Conference on Systems, Man, and Cybernetics (SMC). pp. 3373–3378. doi:10.1109/SMC.2017.8123151.

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39, 2481–2495. doi:10.1109/TPAMI.2016.2644615.

Ballard, D.H., 1981. Generalizing the hough transform to detect arbitrary shapes. Pattern Recognition 13, 111–122. doi:10.1016/0031-3203(81)90009-1.

Bodenstedt, S., Allan, M., Agustinos, A., Du, X., Garcia-Peraza-Herrera, L., Kenngott, H., Kurmann, T., Mller-Stich, B., Ourselin, S., Pakhomov, D., Sznitman, R., Teichmann, M., Thoma, M., Vercauteren, T., Voros, S., Wagner, M., Wochner, P., Maier-Hein, L., Stoyanov, D., Speidel, S., 2018. Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery. CoRR abs/1805.02475.

Bourdel, N., Collins, T., Pizarro, D., Debize, C., Gremeau, A.S., Bartoli, A., Canis, M., 2017. Use of augmented reality in laparoscopic gynecology to visualize myomas. Fertility and sterility 107, 737–739. doi:10.1016/j.fertnstert.2016.12.016.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. doi:10.1023/A:1010933404324.

Cano, A.M., Gaya, F., Lamata, P., Sanchez-Gonzalez, P., Gomez, E.J., 2008. Laparoscopic tool tracking method for augmented reality surgical applications. International Symposium on Biomedical Simulation 5104, 191–196. doi:10.1007/978-3-540-70521-5_21.

Chaurasia, A., Culurciello, E., 2017. LinkNet: Exploiting encoder representations for efficient semantic segmentation, IEEE Visual Communications and Image Processing (VCIP). pp. 1–4. doi:10.1109/VCIP.2017.8305148.

Chen, L., Tang, W., John, N.W., 2017. Real-time geometry-aware augmented reality in minimally invasive surgery. IET Healthcare Technology Letters 4, 163–167. doi:10.1049/htl.2017.0068.

Choi, B., Jo, K., Choi, S., Choi, J., 2017. Surgical-tools detection based on convolutional neural network in laparoscopic robot-assisted surgery, Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 1756–1759. doi:10.1109/EMBC.2017.8037183.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions, IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1800–1807. doi:10.1109/CVPR.2017.195.

Collins, T., Bartoli, A., 2014. Infinitesimal plane-based pose estimation. International Journal of Computer Vision 109, 252–286. doi:10.1007/s11263-014-0725-5.

Davison, A.J., Reid, I.D., Molton, N.D., Stasse, O., 2007. MonoSLAM: Real-time single camera slam. IEEE Transactions on Pattern Analysis and Machine Intelligence 29, 1052–1067. doi:10.1109/TPAMI.2007.1049.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database, IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. doi:10.1109/CVPR.2009.5206848.

Doignon, C., Graebling, P., Mathelin, M., 2005. Real-time segmentation of surgical instruments inside the abdominal cavity using a joint hue saturation color feature. Real-Time Imaging 11, 429–442. doi:10.1016/j.rti.2005.06.008.

Du, X., Clancy, N., Arya, S., Hanna, G.B., Kelly, J., Elson, D.S., Stoyanov, D., 2015. Robust surface tracking combining features, intensity and illumination compensation. International Journal of Computer Assisted Radiology and Surgery 10, 1915–1926. doi:10.1007/s11548-015-1243-9.

Du, X., Kurmann, T., Chang, P., Allan, M., Ourselin, S., Sznitman, R., Kelly, J.D., Stoyanov, D., 2018. Articulated multi-instrument 2-D pose estimation using fully convolutional networks. IEEE Transactions on Medical Imaging 37, 1276–1287. doi:10.1109/TMI.2017.2787672.

Feuerstein, M., Reichl, T., Vogel, J., Schneider, A., Feussner, H., Navab, N., 2007. Magneto-optic tracking of a flexible laparoscopic ultrasound transducer for laparoscope augmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention 4791, 458–466. doi:10.1007/978-3-540-75757-3_56.

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Martinez-Gonzalez, P., Garcia-Rodriguez, J., 2018. A survey on deep learning techniques for image and video semantic segmentation. Applied Soft Computing 70, 41–65. doi:10.1016/j.asoc.2018.05.018.

Garcia-Peraza-Herrera, L.C., Li, W., Fidon, L., Gruijthuijsen, C., Devreker, A., Attilakos, G., Deprest, J., Poorten, E.V., Stoyanov, D., Vercauteren, T., Ourselin, S., 2017a. ToolNet: Holistically-nested real-time segmentation of robotic surgical tools, International Conference on Intelligent Robots and Systems (IROS). doi:10.1109/IROS.2017.8206462.

Garcia-Peraza-Herrera, L.C., Li, W., Gruijthuijsen, C., Devreker, A., Attilakos, G., Deprest, J., Poorten, E.V., Stoyanov, D., Vercauteren, T., Ourselin, S., 2017b. Real-time segmentation of non-rigid surgical tools based on deep learning and tracking. Computer-Assisted and Robotic Endoscopy. CARE 2016. Lecture Notes in Computer Science 10170, 84–95. doi:10.1007/978-3-319-54057-3_8.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778. doi:10.1109/CVPR.2016.90.

Iglovikov, V., Shvets, A., 2018. TernausNet: U-Net with VGG-11 encoder pre-trained on ImageNet for image segmentation. CoRR abs/1801.05746.

Jaffray, B., 2005. Minimally invasive surgery. Arch Dis Child. 90, 537–542. doi:10.1136/adc.2004.062760.

Jayarathne, U.L., McLeod, A.J., Peters, T.M., Chen, E.C.S., 2013. Robust intraoperative us probe tracking using a monocular endoscopic camera. International Conference on Medical Image Computing and Computer-Assisted Intervention 8151, 363–370. doi:10.1007/978-3-642-40760-4_46.

Jin, A., Yeung, S., Jopling, J., Krause, J., Azagury, D., Milstein, A., Fei-Fei, L., 2018. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks, IEEE Winter Conference on Applications of Computer Vision (WACV). doi:10.1109/wacv.2018.00081.

Kaiser, L., Gomez, A.N., Shazeer, N., Vaswani, A., Parmar, N., Jones, L., Uszkoreit, J., 2017. One model to learn them all. arXiv:1706.05137v1 .

Kim, J., Bartoli, A., Collins, T., Hartley, R., 2012. Tracking by detection for interactive image augmentation in laparoscopy. Lecture Notes in Computer Science 7359, 246–255. doi:10.1007/978-3-642-31340-0_26.

Krupa, A., Gangloff, J., Doignon, C., Mathelin, M.F., Morel, G., Leroy, J., Soler, L., Marescaux, J., 2003. Autonomous 3-D positioning of surgical instruments in robotized laparoscopic surgery using visual servoing. IEEE Transactions on Robotics and Automation 19, 842–853. doi:10.1109/TRA.2003.817086.

Laina, I., Rieke, N., Rupprecht, C., Vizcaino, J.P., Eslami, A., Tombari, F., Navab, N., 2017. Concurrent segmentation and localization for tracking of surgical instruments. International Conference on Medical Image Computing and Computer-Assisted Intervention 10434, 664–672. doi:10.1007/978-3-319-66185-8_75.

Li, Y., Chen, C., Huang, X., Huang, J., 2014. Instrument tracking via online learning in retinal microsurgery. International Conference on Medical Image Computing and Computer Assisted Interventions 8673, 464–471. doi:10.1007/978-3-319-10404-1_58.

Lin, M., Chen, Q., Yan, S., 2013. Network in network, International Conference on Learning Representations (ICLR). doi:arXiv:1312.4400.

MICCAI, 2015. EndoVis challenge. 2015, URL: https://endovis.grand-challenge.org/.

Mottaghi, R., Chen, X., Liu, X., Cho, N., Lee, S., Fidler, S., Urtasun, R., Yuille, A., 2014. The role of context for object detection and semantic segmentation in the wild, IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 891–898. doi:10.13140/2.1.2577.6000.

Odena, A., Dumoulin, V., Olah, C., 2016. Deconvolution and checkerboard artifacts, Distill. doi:10.23915/distill.00003.

Ozgur, E., Lafont, A., Bartoli, A., 2017. Visualizing in-organ tumors in augmented monocular laparoscopy, IEEE International Symposium on Mixed and Augmented Reality (ISMAR). pp. 46–51. doi:10.1109/ISMAR-Adjunct.2017.30.

Pakhomov, D., Premachandran, V., Allan, M., Azizian, M., Navab, N., 2017. Deep residual learning for instrument segmentation in robotic surgery. CoRR abs/1703.08580.

Pezzementi, Z., Voros, S., Hager, G.D., 2009. Articulated object tracking by rendering consistent appearance parts, IEEE International Conference on Robotics and Automation. doi:10.1109/ROBOT.2009.5152374.

Pratt, P., Jaeger, A., Hughes-Hallett, A., Mayer, E., Vale, J., Darzi, A., Peters, T., Yang, G.Z., 2015. Robust ultrasound probe tracking: initial clinical experiences during robot-assisted partial nephrec-

tomy. International Journal of Computer Assisted Radiology and Surgery 10, 1905–1913. doi:10.1007/s11548-015-1279-x.

Puerto-Souza, G.A., Mariottini, G.L., 2013. Toward long-term and accurate augmented-reality display for minimally-invasive surgery, IEEE International Conference on Robotics and Automation. pp. 5384–5389. doi:10.1109/ICRA.2013.6631349.

Rasmus, A., Valpola, H., Honkala, M., Berglund, M., Raiko, T., 2015. Semi-supervised learning with ladder networks. 28th International Conference on Neural Information Processing Systems 2, 3546–3554.

Reiter, A., Allen, P.K., Zhao, T., 2012. Feature classification for tracking articulated surgical tools. International Conference on Medical Image Computing and Computer Assisted Interventions 7511, 592–600. doi:10.1007/978-3-642-33418-4_73.

Rieke, N., Tan, D.J., Alsheakhali, M., Tombari, F., Filippo, C.A.S., Belagiannis, V., Eslami, A., Navab, N., 2015. Surgical tool tracking and pose estimation in retinal microsurgery. International Conference on Medical Image Computing and Computer Assisted Interventions 9349, 266–273. doi:10.1007/978-3-319-24553-9_33.

Rieke, N., Tan, D.J., Tombari, F., Vizcaino, J.P., d. S. Filippo, C.A., Eslami, A., Navab, N., 2016. Real-time online adaption for robust instrument tracking and pose estimation. International Conference on Medical Image Computing and Computer-Assisted Intervention 9900, 422–430. doi:10.1007/978-3-319-46720-7_49.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention 9351, 234–241. doi:10.1007/978-3-319-24574-4_28.

Salah, Z., Preim, B., Elolf, E., Franke, J., Rose, G., 2011. Improved navigated spine surgery utilizing augmented reality visualization, Bildverarbeitung fr die Medizin. pp. 319–323. doi:10.1007/978-3-642-19335-4_66.

Shan, J., Cheng, H.D., Wang, Y., 2008. A novel automatic seed point selection algorithm for breast ultrasound images, 19th Internation Conference on Pattern Recognition. pp. 1–4. doi:10.1109/ICPR.2008.4761336.

Shelhamer, E., Long, J., Darrell, T., 2017. Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39, 640–651. doi:10.1109/TPAMI.2016.2572683.

Shvets, A., Rakhlin, A., Kalinin, A.A., Iglovikov, V., 2018. Automatic instrument segmentation in robot-assisted surgery using deep learning, IEEE International Conference on Machine Learning and Applications (ICMLA). doi:10.1109/ICMLA.2018.00100.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 .

Smith, L.N., 2015. Cyclical learning rates for training neural networks. arXiv:1506.01186 .

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research 15, 1929–1958.

Suzuki, S., Be, K., 1985. Topological structural analysis of digitized binary images by border following. Computer Vision, Graphics, and Image Processing 30, 32–46. doi:10.1016/0734-189X(85)90016-7.

Wang, M., Wu, J., Lee, P., Hu, M., Kumar, A., Chen, L., Liu, K., Marescaux, J., Nicolau, S., Suraj, A., Vemuri, Soler, L., 2012. A landmark based registration technique for minimally invasive spinal surgery, IEEE International Symposium on Consumer Electronics (ISCE). pp. 235–236. doi:10.1109/ISCE.2013.6570203.

Wei, G., Arbter, K., Hirzinger, G., 1997. Automatic tracking of laparoscopic instruments by color coding. International Conference on Computer Vision, Virtual Reality, and Robotics in Medicine 1205, 357–366. doi:10.1007/BFb0029257.

Wengert, C., Bossard, L., Haberling, A.H., Baur, C., Szekely, G., Cattin, P.C., 2008. Endoscopic navigation for minimally invasive suturing. Comput Aided Surg 13, 299–310. doi:10.3109/10929080802337914.

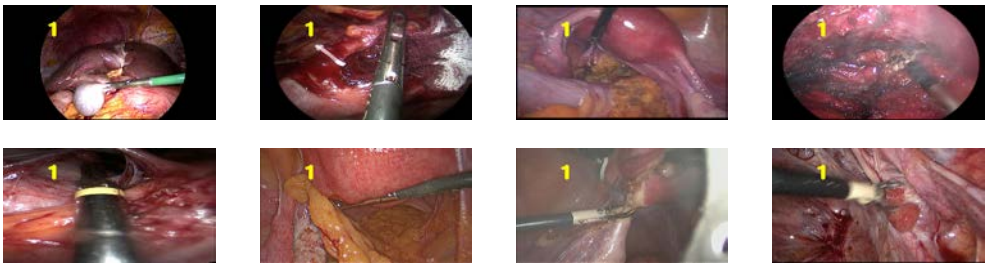
Wu, Y., Hu, Z., 2006. PnP problem revisited. Journal of Mathematical Imaging and Vision 24, 131–141. doi:10.1007/s10851-005-3617-z.

Xie, S., Tu, Z., 2015. Holistically-nested edge detection. arXiv:1504.06375 .

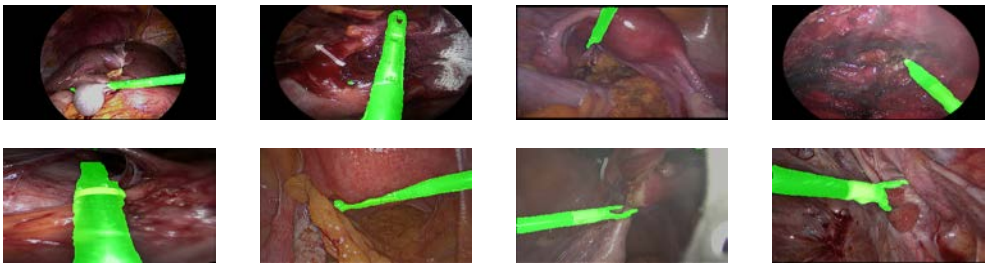
Yuan, Y., Lo, Y., 2019. Improving dermoscopic image segmentation with enhanced convolutional-deconvolutional networks. IEEE Journal of Biomedical and Health Informatics 23, 519–526. doi:10.1109/JBHI.2017.2787487.

Zeiler, M.D., 2012. Adadelta: an adaptive learning rate method. CoRR abs/1212.5701.

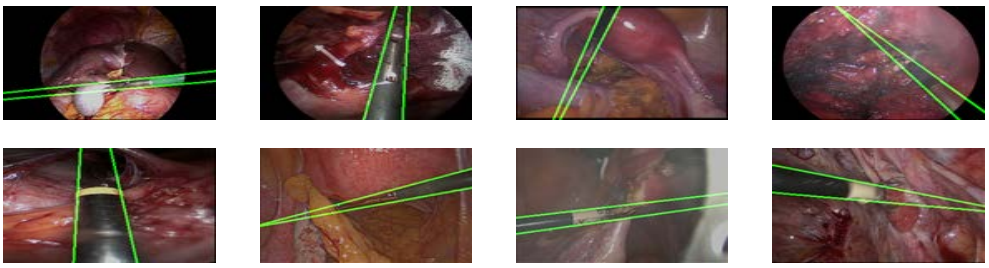
Appendix A. Detection, segmentation, and features extraction results of the noisy laparoscopic images as shown in Fig. 1 from our proposed ART-Net.



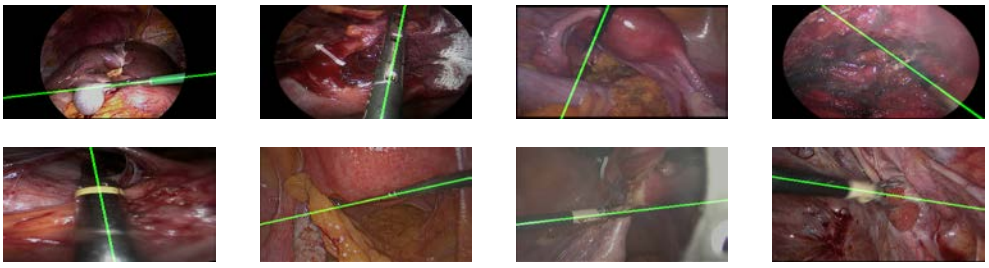
Detection results of surgical instrument where tag "1" indicates tool presence



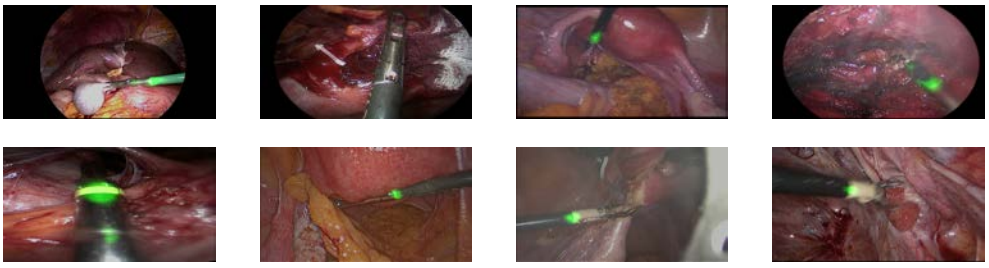
Segmentation results of surgical instrument



Edge-line (tool boundaries) extraction of surgical instrument



Mid-line extraction of surgical instrument



Tip-point extraction of surgical instrument

Medical Imaging and Applications

Master Thesis, June 2019

Direct 3D printing from DICOM images

María Natalia Herrera Murillo, **Supervisors**: Dr. Oliver Diaz Dr. Robert Martí

Computer Vision and Robotics Group, University of Girona, Catalonia, Spain

Abstract

Over the past years, 3D printing technologies have been used to convert medical images into models of specific organs of the human body. These models allow physicians to get more interaction and an improved visual perception of the patients anatomy. Typically, these models are printed from an isosurface, previously saved as a STL, VRML, 3MF, VTK u other file formats. However, this traditional approach is centered on printing with homogeneous materials and colors for the whole model of the organ. To circumvent this limitation, voxel-based printing offers novel features to print volumes containing material properties at the voxel level, opening the door for the production of biomechanically accurate models in the future. In this work, a voxel-based framework for 3D printing is proposed, based on both MRI and Magnetic Resonance Elastography (MRE). The combination of these modalities intended to allow the mapping of the intensities to the corresponding stiffness of the tissues. The framework consists of the following five steps. First, the pre-processing of the images, using bias field correction and Principal Component Analysis (PCA), when required by the images. Then, the segmentation of the images to extract the organ of interest. Among many segmentation techniques available and multiple approaches to implement them, this project took advantage of a semi-automatic segmentation based on a global thresholding, morphological operations and region properties. This segmentation was applied in the case of images having an appropriate level of contrast to differentiate the organ. Otherwise, in the case of images having more challenging characteristics for the segmentation, a manual approach was taken. Once the segmentation was completed, the stiffness values were assigned to one section of the organ by means of the elastography images. The finals steps were related to the preparation of the slices in a format compatible with the settings of the printer.

Keywords: 3D printing, additive manufacturing, voxel printing, bitmap, Magnetic Resonance Elastography, stiffness

1. Introduction

3D printing, known as additive manufacturing or rapid prototyping (Chang et al., 2019), is a recent strategy used in hospitals to produce realistic models of an organ. Such printed organs represent a huge innovation for the medical field in terms of customization, prototyping, manufacturing, and research. Besides, this technology has created the opportunity to directly interact with a model of a patient-specific anatomy, allowing surgeons to reach a superior 3D perception and tactile feedback. In general, 3D printing has the potential to add more accuracy and effectiveness in the operational process.

There is a broad range of applications of 3D printing in the medical domain. Just to mention a few: 3D phantoms mimicking the breast have been designed to replicate the tissue attenuation profile given by real 2D mammograms and evaluate the performance of mammography systems (Clark et al., 2016), cardiac models have been used to simulate anatomic and functional properties of aortic valve stenosis (Maragiannis et al., 2015) and surgical planning of device placement in patients with congenital cardiac diseases (Farooqi et al., 2016), kidney models with tumors have been created to study minimally invasive surgery procedures (Zhang et al., 2016), while models of the brain are used to reproduce the soft nature of this organ and mimic its tactile properties (Ploch et al., 2016). Going even further, an extension of traditional 3D printing known as bioprint-

ing or tissue fabrication studies different approaches that incorporate biological products from materials such as living cells (Agarwala, 2016).

The versatility of 3D printing is accelerating and transforming different areas, particularly in the medical domain. Its potential guarantees that this technology will continue growing and shaping the future of the the health care industry for the years and decades to come.

1.1. 3D Printing

1.1.1. Typical work-flow of 3D printing

Traditional approaches to develop a 3D printed model could be simplified into 5 major steps (see Figure 1). The first step is the imaging of the anatomy of interest by means of multiple modalities that produce volumetric datasets, such as Computer Tomography (CT), 3D Angiography and Magnetic Resonance Imaging (MRI). The images obtained through these modalities are primarily stored in a DICOM format. The second step is the segmentation of the organ or tissue, creating its corresponding mask. This step heavily depends on the high spatial and contrast resolution. Segmentation methods include thresholding, watershed, region growing, edge based detection, and so on. The third step is the generation of a Computer-Aided Design (CAD) 3D model directly based on that mask. Next, the model is converted to the appropriate printing format and it is finally printed. Selecting the most suitable combination of algorithms, techniques, and tools at each of the aforementioned levels could expedite the process, and most importantly, provide more reliable results.

1.1.2. 3D Printing techniques

Concerning 3D printing techniques in the clinical setting, it is possible to distinguish between different processes and materials available. Some techniques melt or soften the materials which are then disposed in a layer-by-layer fashion, such as selective laser sintering (SLS) or fused deposition modeling (FDM). Other more sophisticated and expensive techniques cure materials by a light source, that is the case of stereolithography (SLA) and multijet modeling (MJM) (Chae et al., 2015; Chang et al., 2019). The last two techniques print products of higher resolution, so the models have a smoother surface finish and internal structures with better accuracy. To select the most suitable technique, the main considerations to ponder include the cost of the printer, the cost of the materials, speed, resolution, surface finishing, post-production and handling of multiple colors and materials.

Currently, FDM printers are the preferred option in medicine because of their affordability, acceptable accuracy, minimal maintenance and high availability of printers in the market (Chae et al., 2015). On the other hand, printers based on this technique have limitations on the number of colors and materials handled. To help mitigate this issue, one approach consists of printing multiple nested 3D model files, however, this process can be labor intensive and very inefficient.

1.1.3. Software packages for 3D Reconstructions and Mesh Repairing

To fabricate a 3D model, basically, modeling software is used to segment the intended organ or tissue. Then, the segmented region is exported as a 3D model, usually in STL format. Other formats include VRML, 3MF, VTK and PLY, for example. Once a model is created, it can be optimized by many options such as editing, cleaning, healing, inspecting, rendering, texturing and, in general, preparation for printing. Nevertheless, it is important to note that most of the 3D reconstructions solutions lack post-processing tools (Chang et al., 2019). This situation is compensated by software specifically used for mesh repairing, available from multiple companies. For more information, in Appendix I, a list of software packages for 3D reconstructions and mesh repairing is provided.

Additionally, there are 3D slicing software packages that divide the CAD file into thin data slices suitable for the printer settings. However, slicing is usually performed by the proprietary software accompanying the printers.

1.2. Voxel Printing

Traditional approaches are centered on printing with homogeneous materials and colors for the whole model or sections of it, but 3D printing has been evolving significantly. Nowadays, several printers have novel features to print 3D volumes containing material properties at the voxel level. Voxel printing, also described as bitmap-based printing (Doubrovski et al., 2015), allows defining a color for each individual voxel and print models with highly complex material and color distribution. Just to obtain an idea about the potential of this approach: “With voxel printing that has the capability of a trillion voxels in the space of the printer and six different materials, the number of possible combinations is six to the power of a trillion, which is an astronomical number” (Stratasys, 2017).

In contrast to the traditional 3D model generation (see Figure 1), which only uses the binary mask to create the model, voxel printing applies the mask to the organ of interest to extract and preserve the intensity gradients of the native image (Hosny et al., 2018; Solomon et al., 2016). The masked data sets are then re-slice into the printer’s native X, Y and Z resolutions. Besides,

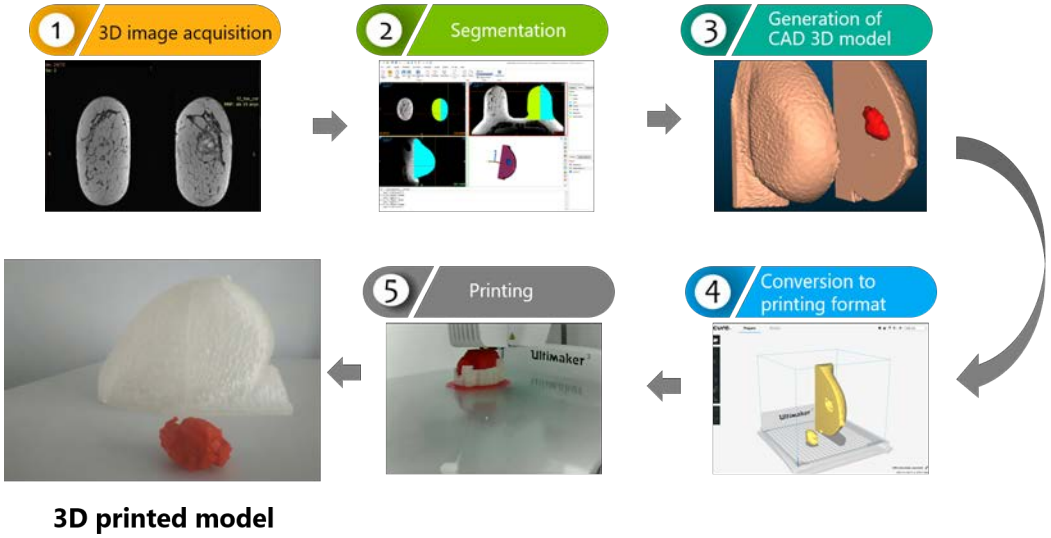


Figure 1: Typical work-flow. This breast phantom was created in the Hospital Parc Taulí, using Materialise software for the creation of the CAD model, MeshLab for optimizing the STL model and Ultimaker 3 printer.

it is important to highlight that this method avoids the generation of a 3D model (in most of the cases). In any case, bitmap slices such as BMP, GIF, JPEG, EXIF, PNG, and TIFF files, for example, are the direct input of the printer.

Bitmaps are defined as a regular rectangular mesh of cells called pixels, each of the pixels containing a color value. Moreover, bitmaps are characterized by only two parameters, the information content per pixel and the number of pixels. There are other parameters that are applied to them but they are derivations of these two principal parameters. They are always orientated horizontally and vertically. For bitmap generation, pixels should be considered square although they may have other aspect ratios in practice.

The principal idea about the use of bitmaps is that in volumetric scans each voxel is represented by a range of shades. Whilst the generated organ bitmaps are binary and feed the 3D printer with information about whether the material has to be placed on the location of the voxel or not.

Bitmap printing is used for printing high resolution 3D organs and to generate smooth and uniform transitions between materials of different hardness. Also, it avoids sharp edges of transition between hard and soft materials, where files to the printer are written on a local, voxel-scale level.

1.3. Objective

The main aim of this work is to develop a framework for voxel-based printing, using both MRI and magnetic resonance elastography (MRE) images. The combination of these modalities intends to allow the mapping of the intensities to the corresponding stiffness of the tissues and, in the near future, to facilitate assigning materials based on desired graded mechanical properties. To accomplish the objective of this project, some approaches concerning segmentation, mapping of elastic properties and slicing of the models were explored.

This project has been developed in collaboration with the Digital Medical Imaging Center (CIMD) at Hospital Parc Taulí and the company AVINENT, both in Barcelona. The hospital counts on an Ultimaker 3 printer, built for FDM, and used with Cura Ultimaker to prepare the printer settings. Since this printer contains a dual extrusion system, it allows a dual combination of build and water-soluble support materials. As an example of a recent work performed in the hospital, Figure 2 shows the printing of an aneurysm.

In the case of AVINENT, a printer Stratasys J750, which is based on an MJM technology, allows to print at voxel level and create models that fully exploit multi-material and multi-color capabilities. The printer Stratasys J750 allows to load up to six materials at once, and print with a resolution as fine as 0.014 mm (Stratasys, 2019). The desired outcome is to create the correct input bitmaps for the printer. So in the future, it could be possible to print more realistic organs, using this advanced equipment and new materials.

1.4. Added value of MRE to MRI

MRE is a non-invasive technique for imaging the mechanical properties of soft tissues, which has an acquisition time of about one minute and is acquired as an add-on to the standard MRI exam (Low et al.,

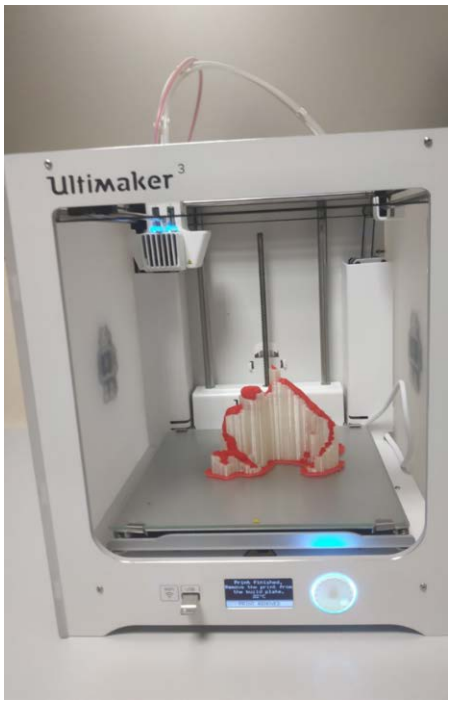


Figure 2: Printing of an aneurysm in Hospital Parc Taulí, using Seg3D software for the creation of the CAD model, MeshLab for optimizing the STL model and Ultimaker 3 printer. Aneurysm is printed in red whereas the support material is in white.

2016; Venkatesh et al., 2018). To generate this imaging modality, shear waves are created by applying low-frequency continuous vibrations in specific parts of the surface of the body. These waves are tracked as they pass through various structures, moving more rapidly in stiff material than in soft one. Since MRE measures the waves propagation speed, the obtained images show the wavelength displacement (in μm units), which indirectly reflects the stiffness of such structures (See Figure 3). Then, the wave image is mathematically processed into quantitative images directly showing the tissue stiffness (in kPa units), known as elastograms (Venkatesh et al., 2018).

Currently, the main application of MRE is assessing liver stiffness as an indicator of possible liver fibrosis (Hawley et al., 2017). It is used to complement the anatomical information provided by the MRI because many illnesses do not cause significant anatomical changes until they reach an advanced state. Instead, they can affect the stiffness of the tissues in a profound way. MRE has also been proposed to identify diseases in other organs: for example in the breast to detect cancer, fibroadenoma, fibrocystic; in the uterus to detect fibroids; in the pancreas to detect pancreatitis or adenocarcinoma, in the spleen to detect portal hypertension or even in the brain to differentiate between hard and soft meningioma (Hawley et al., 2017; Mariappan et al., 2010; Murphy et al., 2013).

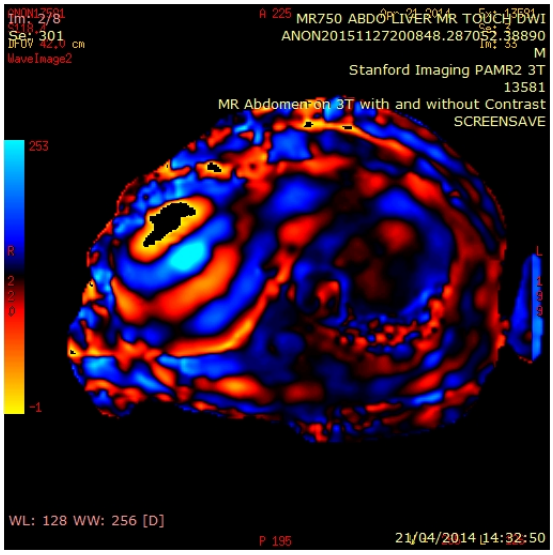


Figure 3: Wave image of a section of the liver, in which the color coding represents the displacements caused by the propagating waves, in μm. This image was provided by Stanford Medicine Imaging Center (Stanford, 2019).

2. State of the art

In this section, we will review some of the most recent and relevant voxel-based printing methods published in research journals. The subdivision of the literature is also based on the similarity of the approaches implemented by the authors.

In 2015, Doubrovski et al. proposed a printing process applied to the design of a customized prosthetic socket. Creating the sockets using a multi-material 3D printing allows locally varying material stiffness and increases its functionality. In this model, the geometry of the object along with its desired material properties is determined. The collection of the data is based on three sources: MRI scans of the residual limb of a patient, 3D scan data of the surface of an existing socket and the pressure distribution along with the socket-limb interface.

Then, bitmaps are used to map the desired properties into local material composition, matching the resolution of the printer. Similarly, other approaches create high-resolution, geometrically complex and materially heterogeneous 3D objects (Bader et al., 2016, 2018; Weeger et al., 2016). In the case of Bader et al, the authors implemented a procedure called Titled Data-driven Material Modeling (DdMM) to generate slices having heterogeneous material distributions. The previous condition is achieved by integrating multiple geometry-based data sources, such as point-clouds, scalar and vector fields, curves and polygons, and tetrahedral meshes. Furthermore, this framework uses polygonal meshes which are later rasterized in layers.

Something particular about these layers is their capacity of storing material information, vectors for velocity fields, matrices and other data structures in each voxel. Finally, dithering is used to create binary bitmaps, used to provide instructions for the material deposition during the printing. In the same line of research, an article called "Making data matter: Voxel printing for the digital

fabrication of data across scales and domains" explores voxel-based printing to close the gap between digital information representation and physical material composition (Bader et al., 2018). As in the previous cases, multiple datasets are converted to dithered material deposition descriptions during a layering process. The previous process is performed such that each pixel defines the material identity of a droplet and its placement in 3D space. The printer is able to take these descriptions of the voxels to fabricate the 3D model.

Also in 2015, a 3D printing method based on dithering by an Error Diffusion algorithm was published (Brunton et al., 2015). In this article, the input of the pipeline is a shape defined by a surface with color information. First, the previously mentioned algorithm is used to reproduce the object's albedo texture, that is, a map that defines the color of diffused light. Second, a voxelization procedure is performed, in which a grid of voxels is created according to the specifications of the printer. Also, during the voxelization process, colors are assigned to the surface voxels. The final outcome of this paper's implementation is a realistic reproduction of colors, color gradients and fine details of the initial input.

A method to develop textured 3D phantoms, based on a voxel-based input method, was introduced in 2016 by Solomon et al. In this case, volumes of a liver are segmented from abdominal CT scans. From these scans, a model of CLB textures is extracted, reflecting the real liver textures. For this purpose, a fitting technique called 3D Clustered Lumpy Background (CLB) is applied. At the end, the designed phantoms are voxelized and input into the printer. It is worth mentioning that the custom 3D printing software selected for this implementation included principles of digital dithering by itself.

A less complex framework was developed by Hosny et al., in 2018, for printing high resolution functionally graded multi-material models. To start, the proposed work-flow generates the models from cross-sectional slices of CT and MRI scans. Then, it applies a thresholding-free approach to delimit the organ and generates binary bitmap slices for the printer. The binary bitmaps are the result of the implementation of an algorithm called Floyd–Steinberg dithering. As in some of the previous works, mentioned in this section, a contribution of voxel printing is the preservation of more fine details and information about bio-mechanical gradients (opacity and elasticity gradients). This method bypasses the 3D Model generation and its corresponding slicing.

The authors also show the application of the method for printing structural details in a myocardial infarction model, in order to improve the visualization of scar geometry and surgical planning.

3. Material and methods

3.1. Description of Image Dataset

Three main datasets, corresponding to three different patients, were used in this study. For each patient, there is an anatomical image and its matching elastography images (with and without mask). The imaging area corresponds to the abdomen. MRE is an imaging modality still not widely available in hospitals and obtaining these images is a demanding task. For this project, all the images were provided by Stanford Medicine Imaging Center (Stanford, 2019) and generated with Advantage Workstation 4.6 of GE Medical Systems. Some of the characteristics of the datasets are shown in the tables below:

Patient 1			
Exam: MR Abdomen on 3T, Series: Ax Liver Elast			
DESCRIPTION	A) Volume dimensions	B) Voxel dimensions (mm)	C) Spacing between slices (mm)
Ax T2 RT ABD	512 X 515 X 31	0.8203 X 0.8203 X 8	9
Elastogram	256 X 256 X 8	1.6406 X 1.6406 X 7	9.5
Elastogram + Mask	256 X 256 X 8	1.6406 X 1.6406 X 7	9.5

Table 1: Characteristics of the datasets of images of patient 1.

Patient 2			
Exam: MR Abdomen on 3T, Series: Ax Liver Elast			
DESCRIPTION	A) Volume dimensions	B) Voxel dimensions (mm)	C) Spacing between slices (mm)
Fat Fraction: Ax Liver ME BH	256 X 256 X 16	0.7188 X 0.7188 X 10	10
Elastogram	256 X 256 X 8	1.6406 X 1.6406 X 7	9.5
Elastogram + Mask	256 X 256 X 8	1.6406 X 1.6406 X 7	9.5

Table 2: Characteristics of the datasets of images of patient 2.

Patient 3			
Exam: EOB, Series: 3 plane loc BH			
DESCRIPTION	A) Volume dimensions	B) Voxel dimensions (mm)	C) Spacing between slices (mm)
Water: Ax LAVA-FLEX	512 X 515 X 68	0.6836 X 0.6836 X 6	3
Elastogram	256 X 256 X 4	1.6406 X 1.6406 X 10	10
Elastogram + Mask	256 X 256 X 4	1.6406 X 1.6406 X 10	10

Table 3: Characteristics of the datasets of images of patient 3.

3.2. Software tools

The implementation of the proposed framework was mostly performed on MATLAB 2018b environment. Besides, RadiAnt Dicom Viewer (Medixant, 2019) and 3D Slicer software (HarvardMedicalSchool, 2019) were also used. The first one contributed to the visualization of the images and supported the segmentation tasks, while 3D Slicer was used to perform pre-processing tasks.

3.3. Method

In this section, a framework for voxel-based printing, using both MRI and MRE images, is presented in five main steps (see Figure 4). Each step is explained in detail in the following sub-sections.

3.3.1. Pre-processing

Prior to the segmentation procedure, it is important to consider different factors that could negatively affect the organ mask generation, hence should be corrected. Noise, ringing artifacts or even intensity inhomogeneities are among those factors that can corrupt the images.

Since bias field signals commonly corrupt MRI images (Juntu et al., 2005), a bias field correction was the first method applied to the anatomical images Water: LAVA-FLEX. For this purpose, the N4ITK functionality of 3D Slicer software was used. In this case, as the input parameter is given the volume that potentially could present the in-homogeneity, then a BSpline grid resolution of 1,1,1 is selected, which results in a 4x4x4 grid of control points. Finally, the bias field corrected and bias field volumes are the output of this function (Slicer, 2013).

Second, Principal Component Analysis (PCA) method was computed to reduce noise content and suppress any ringing artifact. This algorithm is a mathematical procedure that uses an orthogonal transformation to convert a set of possibly correlated variables into a set of linearly uncorrelated variables, which are called the principal components (Gonzalez et al., 2002). The applied transformation is done in such a way that the first of the principal components corresponds to the most dominant feature, presenting the largest level of variation. On the other hand, the later components are dominated by undesired elements such as noise and could be removed without any great loss. For this project implementation, a threshold value of 99% was defined to separate the significant components from the ones that are not, as done by Omer et al., 2018. Based on the new variables, a new volume is reconstructed. Such implementation is described in Algorithm 1. Also, it should be mentioned that the algorithm was applied to the whole volume simultaneously, not to the individual slices. So, the input of the algorithm is a grayscale volume, f(m,n,c), of dimensions M x N x C voxels (where M represents the height, N is the width and C is the number of slices).

Algorithm 1 PCA in Matlab

1: ***Data preparation****: standardization of the data*
2: ***Reshape the volume:*** *into a 2D matrix X of dimensions M x NC*
3: ***Translation:*** *Translate the data X to be centered at (0,0).*
4: *Calculate the mean of all the variables:* $\bar{X}$
5: *Translate mean to the origin, by subtracting mean:* $X - \bar{X}$*, from each variable*
6: ***Calculate the covariance matrix*** $\sum$*: for all the variables in the dataset*
7: ***Calculate the eigenvectors and the eigenvalues of the covariance matrix***
8: *From the covariance matrix, compute the eigenvalues* λ_n *and the eigenvectors that go along with them. Each eigenvalue gives the variance of the data in the direction of the correspondent eigenvector.*
9: *Sort eigenvalues and eigenvectors in descending order, following these criteria:* $\lambda_1 > \lambda_2 > \lambda_3 > ...\lambda_n$
10: *From the eigenvalues, compute the total variance:* $V = \sum_{j=1}^{n} \lambda_j$.
11: ***Choose principal components***
12: *Choose the k largest eigenvalues (eigenvectors) to account for the desired percentage of variation:* $\sum_{j=1}^{k} \lambda_k / V >= 99\%$*, for example.*
13: ***Compute the reconstructed dataset:*** *project the input vectors by multiplying the data set by a matrix of the reduced eigenvectors A, (where the columns* λ_n *are the eigenvectors corresponding to the largest k eigenvalues), new data = AX*

3.3.2. Organ segmentation

The axial slices of the elastography were imaged in the widest portion of the liver, for this reason, the next step was the extraction of the liver mask. This step represented a big challenge because the level of contrast between the liver, surrounding organs, and abdominal fat was different among the anatomical images of the three patients. In the T2 images, the hepatic tissue is presented darker than (hypointense to) the spleen and kidneys but there is not enough contrast of the liver

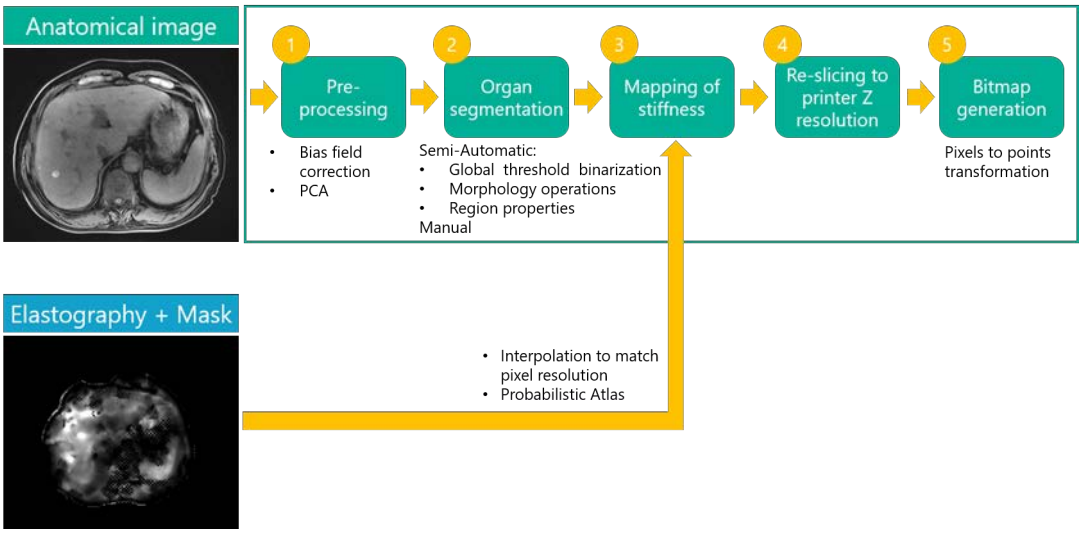


Figure 4: Framework for voxel-based printing develop during this thesis project.

with respect to the abdominal fat. In the case of Fat Fraction, the images display a quantitative map of the percentage of hepatic fat signal which is not adequate for the segmentation task. Conversely, Water: LAVA-FLEX images are generated by using a technique that provides homogeneous fat suppression over the entire image (GE, 2018). The Water images presented an adequate level of contrast with respect to surrounding tissues.

Given the previous reasons, a hand segmentation approach was applied for the T2 and Fat Fraction volumes. Basically, the contour of the hepatic tissue was drawn manually, for all the slices, using RadiAnt Dicom Viewer drawing tools. Then, these images were exported to Matlab, where the contour was detected by subtracting the red from the green channel of the images and applying a threshold. Once the contour was detected, the area inside it was filled. Because the contours were drawn manually, some subjectivity is unavoidable.

For the Water image, a semi-automatic segmentation was chosen. First, a global threshold binarization was used to isolate the tissue of the main organs in the abdomen and get rid of the structures in the background. The computed thresholding algorithm is explained in Algorithm 2. Basically, this algorithm initializes a threshold as the average value of adding the highest intensity to the lowest intensity of the image. Then, it uses this threshold to separate the variables in two groups of intensities. From both groups, the average intensity is calculated. The threshold is re-calculated, but this time, as the average of the average intensities of the two groups. Except for the initialization, the remaining operations are repeating until reaching a satisfactory threshold.

To improve the result of the global thresholding, the

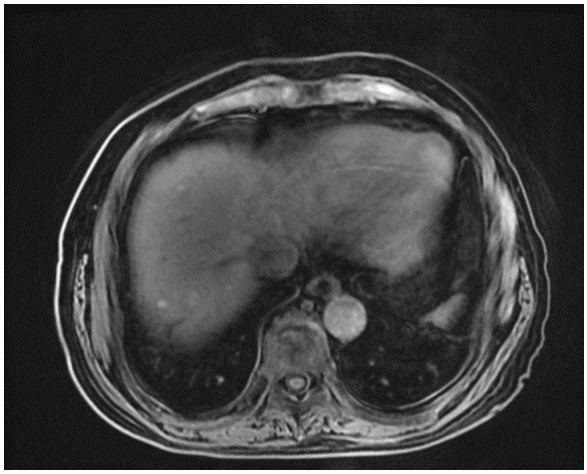


Figure 5: Water: LAVA-FLEX image. In this image, it is appreciated how the intensities of the liver and the stomach are fused. This image was provided by Stanford Medicine Imaging Center (Stanford, 2019).

Matlab function 'imfill' is used to fill regions inside the mask. A morphology operation opening, dilation of the erosion, is also implemented to remove small objects from the foreground. This last step intended to improve the contour of the organs, especially in some images in which the tissue of the liver is fused with the tissue of other organs, such as kidney and stomach (see Figure 5).

Also, the separation process of these tissues was supported by a graphical user interface (GUI), quickly designed using 'App Designer' of Matlab. 'App Designer' is particularly useful because automatically generates object-oriented code that specifies layout and design (MathWorks, 2019). This GUI relied on the function 'drawfreehand' to select and delete undesired areas of the mask. After using the GUI, an opening is applied again to correct any undesired effect of the manual dele-

Algorithm 2 Global threshold binarization in Matlab (Karthikeyan and Valliammai, 2012)

1: ***Set an initial threshold value:*** *as T = (maximum intensity value of the volume + minimum intensity value of the volume) /2*
2: ***For*** *i = 1 : number of iterations*
3: ***Get two sets of pixels:*** *using T to segment the image. In one set include the pixel values less than T (set1) and include the pixel values higher than T (set2) in the other group.*
4: ***Calculate the average*** *value of each group of pixels*
5: ***Re-calculate threshold*** *as T = (average intensity value of set1 + average intensity value of set2) /2*
6: ***End***
7: ***Create the binary mask*** *according to the calculated threshold T*

tion. Finally, since the function 'regionprops3' of Matlab measures a set of given properties, for each connected component of a volumetric binary image, it was employed to measure the volume of each of these components. The biggest count of the number of 'on' voxels in each region was selected since it corresponds to the liver. So, the final output of this process is the mask of the liver.

3.3.3. Mapping stiffness properties

One of the aims of the proposed framework is to assign stiffness properties to the voxel intensities of the anatomical image. The stiffness properties are obtained from the elastography that includes a mask. The mask or confidence map is important because it crosses out the areas of the liver of unreliable measurement, such areas could include large blood vessels, the Glisson's capsule or any area with wave interference (Tang et al., 2015). Before starting this step, it is also important to considerer that the elastography imaged a wide portion of the liver, not the whole organ (see Figure 6).

In order to directly map the stiffness to the matching section of the anatomical volume, first, it was considered that the two volumes do not have the same pixel resolution or dimensions. So, it was chosen to use a function in Matlab to compute an interpolation for 3-D gridded data in a meshgrid format. The interpolation considered the elastography as the image to be interpolated, and X_e, Y_e and Z_e as its coordinates in each axis. Since the desired resolution is the one of the anatomical volume, its coordinates X_a, Y_a and Z_a were also provided for the interpolation function.

Since the elastography images of patients 1 and 2 (see Tables 1 and 2) have a slice increment bigger than the slice thickness, new information is being created. In the case of patient 3, these parameters are the same so the original slices are just re-sliced. Besides, it was possible to select the interpolation method, so, a 'linear' approach was given as a parameter. The linear interpolation is based on the values at the neighboring grid points, in each of the dimensions.

Once the volumes had the same resolution, and dimensions in X, Y and Z are matched, it is possible to directly know the stiffness values of one section of the liver in the anatomical image. And now, the question is how to obtain the stiffness of the remaining part of the liver. The approach taken to find those values was the generation of a Multi-class Probabilistic Atlas.

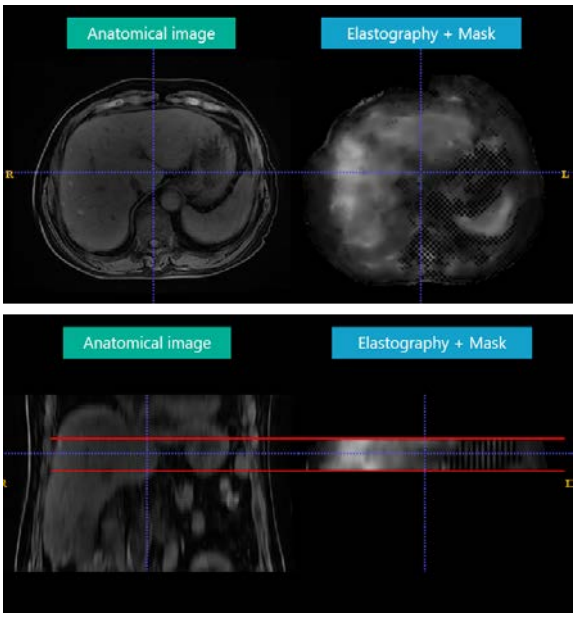


Figure 6: MRI and Elastography slices of patient 3. The four axial slices of the elastogram are prescribed so that the liver is imaged in its widest portion. These images were provided by Stanford Medicine Imaging Center (Stanford, 2019).

A Probabilistic Atlas answers to the problem of how to estimate a label X, that better explains a given observation Y, that is the probability of $P(X|Y)$ (Gubern-Mérida et al., 2011). Based on the theorem of Bayes, such probability can also be written as $P(Y|X)P(X)$.

In this case, the starting point to implement this method was the K-means algorithm of Matlab, which clustered the stiffness values into 4 groups and returned a segmented labeled volume L having the labels 1, 2, 3 and 4. Also, it returns the centroid locations of those clusters. The next task is the computation of the probabilistic atlas, one of them for each of the 4 clusters, showing the frequency with which each voxel was labeled as belonging to such cluster. So, the probability distribution $P(X)$ of cluster 1, 2, 3 and 4 is given by this probabilistic atlas.

On the other hand, since the elastography and the anatomical image are in the same reference space, it was possible to directly build a model showing the distribution of the normalized intensity values, given a particular segmentation X, that is, $P(Y|X)$. This model $P(Y|X)$ is based on the part of the liver where the stiffness values are known. In the other parts of the liver, the estimated segmentation X is obtained based on the maximization of the posterior probability $P(Y|X)P(X)$. In general, this method was considered for this work because it is able of providing both a reference space and spatial distribution of probabilities of a voxel belonging to a class (Gubern-Mérida et al., 2011).

The centroid values given by the K-means algorithm, and associated with each cluster, are assigned to the intensity values in the anatomical image. To conform the final volume of stiffness values, the interpolated elastography and the anatomical image (having the stiffness mapped) are combined. Finally, the volume of stiffness is masked using the liver mask generated from the anatomical images, in the previous subsection of this report.

3.3.4. Re-slicing to printer Z resolution

Every slice in the voxel print must have the same pixel dimensions (width and height). Also, it is necessary to slice the desired final object at a height that matches the printer layer thickness. In the case of the printer J750 of Stratasys, there are two settings: High Quality Mode, that prints 0.014 mm between layers and High Speed/High Mix Modes that uses 0.027 between layers (Community, 2019). If the slice thickness is less than the printer setting, the printer will try to compensate for the difference. For example, if the slices are generated using a thickness of 0.0135 mm and the printer is using a High Speed Mode, it will print each slice twice to reach the desired thickness of 0.027 mm.

The challenges here are the big volume of the liver and the datasets used for this project, which do not have a good pixel resolution (see Tables 1, 2 and 3). Given that Matlab has some limited memory available to allocate in the creation of matrices, it was not possible to re-slice the whole volume at once. Instead, a loop was used to re-slice each original slice, like the one shown in Figure 7, to 0.027 mm. For the previous purpose, the function of interpolation of Matlab was used again, but this time just to re-slice, not creating new information. The output of this step is a matrix of sub-slices for each of the original slices of the volume.

3.3.5. Bitmap generation

The last step of the framework is the generation of bitmaps, in this case as PNG images. For this purpose, the algorithm suggested by (Community, 2019), for the

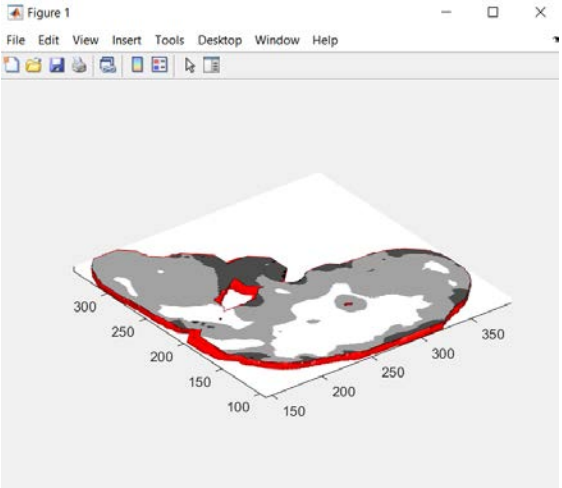


Figure 7: Slice of the liver of patient 3.

printer J750 of Stratasys, was computed for each slice. Basically, instead of saving the images using pixel units, they are saved as points.

There are more complex approaches that imply saving the images as dithered bitmaps, by using algorithms as Floyd–Steinberg (Floyd, 1976) or Error-diffusion (Sullivan et al., 1993). However, this algorithms were not explored during this thesis project.

Algorithm 3 Bitmap generation (Community, 2019)

1: ***Define number of sub-slices***
2: *For index= 1: number of sub-slices*
3: *filenamearray = ["slice_e", index]*
4: *loopfilename = strjoin(filenamearray, '_')*
5: *imshow(figure)*
6: ***Grab the current figure handle:*** *fig=get(groot, 'CurrentFigure')*
7: ***Set the figure units to points:*** *fig.PaperUnits = 'points', as opposed to inches or cm*
8: ***Set the output of the figure to start at location (0,0). To calculate width and height, convert from pixels to points (multiplying by 0.48):*** *fig.PaperPosition = [0 0 wide height]*
9: *axis off*
10: *print(loopfilename, "-dpng")*
11: *End*

4. Results

The implementation was applied in the 3 datasets of images provided for this project. For all the cases, the qualitative and quantitatively results are presented in this section.

In Figure 8 is presented one axial slice, in the widest part of the liver of patient 3. The volume to which this image belong was pre-processed from its original obtained

form. By a visual inspection, it can be noticed that the bias field correction reduced the inhomogeneities in the tissues, and this is specially perceptible in the hepatic tissue. In contrast, the change due to PCA (in terms of reducing noise and artifacts) is not apparent. Nevertheless, the usual effect of this algorithm, that is compression, was obtained; since the amount of components of the volume were significantly reduced. In the bottom row of this same figure, the results of the segmentation are shown. First, global threshold binarization divided the pixels of the image into two groups, allowing to isolate the pixels of the liver and other organs. After applying morphological operations, just three volumes are left the liver, the spleen and the stomach. At the end, based on the size properties of such volumes, it was possible to extract the liver mask.

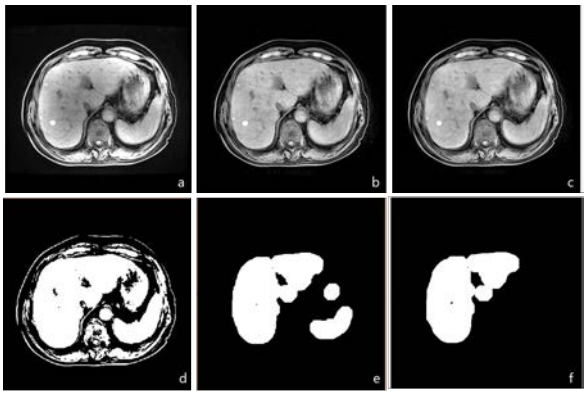


Figure 8: Images of patient 3: (a) Original Image, (b) Image after applying bias field correction, (c) Image after applying PCA, (d) Image after applying global threshold binarization, (e) Image after applying morphological operations and (f) Image after extracting the biggest volume by using region properties.

Figure 9 shows the GUI created using App Designer Tool to delete parts of the stomach that were attached to the liver, negatively interfering in the process of generating the mask of this organ. In this GUI, an 'imfreehand' object encapsulated an interactive free-hand region over the area intended to delete. Among the advantages of this tool is the availability of vertices to adjust the size and position of the polygon by using a mouse. Also, it was possible to drag the polygon over the image, according of the needs of the user.

Another instance of the segmentation results is shown in Figure 10. In this case, axial slices of the widest part of the liver of patients 1 and 2 can be observed. Because of the particular characteristics of these two volumes, it was preferred the manual segmentation, so, the preprocessing steps of bias field correction and PCA were not required. In the first column are displayed the contours drawn using RadiAnt Dicom Viewer. Since the contour drawn in the grayscale image was green, the image was handled in Matlab as an RGB image. The difference between the red and green channels allowed

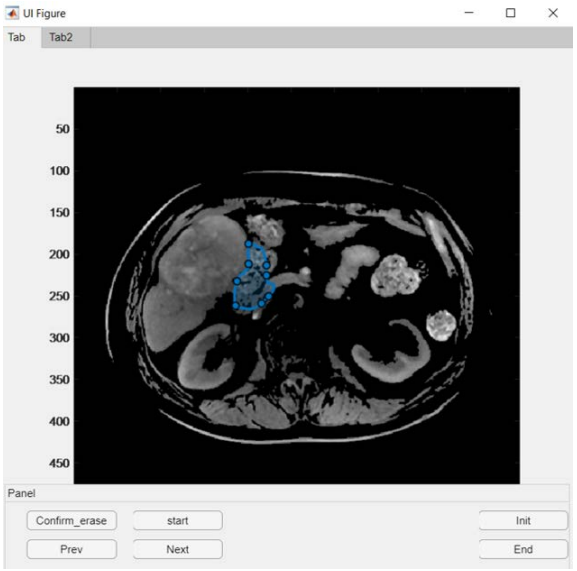


Figure 9: App Designer was used to support the segmentation task.

to isolate the contours, giving the results shown in the middle row of this figure. In those results, it can be observed some discontinuities in the lines of the contour. To address this issue, morphological operations were performed before applying the filling function.

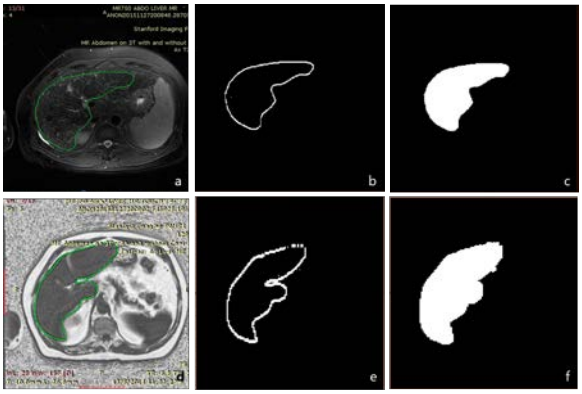


Figure 10: First row: Images of patient 1: (a) Original Image in which the contour was drawn, (b) Isolating the contour of the liver, (c) Filling the contour of the liver; Second row: Images of patient 2.

Now, the results of assigning stiffness values to the liver of the anatomical images are going to be described (see Figure 13); starting by the results of the sections of the liver in which the stiffness was directly obtained from the elastography. Note that in all the cases of the Figure, an anterior view of the liver is displayed. In patient 1, first row, an increased level of stiffness is visualized in the area of the liver, mostly represented by the purple areas. Also in this case, peaks of the stiffness are found in the right posterior and anterior lobes of the liver and in the spleen. For patient 2, the stiffness level stands out in the spleen, where it reaches its maximum level, while in the liver some increment, with respect to the remain-

ing part of the abdomen, can be appreciated. In the case of the third patient, it is clear that the the stiffness is higher in the liver than in any other part of the image. Also for this patient, the peak of the stiffness is in the right lobe of the liver, mainly in the anterior section of it. With regard to the mask of the liver, in patients 1 and 2, it matches the borders of the big mass of increased stiffness in the image. In general, the downside of these results is that once the elastography is masked, there are still pixels in which the stiffness was not assigned. These situation is caused because the measurement obtained in those specific points is not reliable according to the confidence map, generated by the imaging software of GE. This situation affects the results of the liver of patient 2, to a greater extend.

Given the framework adopted in this project, a Probabilistic Atlas approach was chosen to estimate the stiffness in the parts of the liver where this property was not available. The main elements of this atlas are illustrated in Figure 11. To initialize this implementation, k-means was used to create four clusters. This algorithm works by dividing the pixels into clusters, in which each pixel belongs to the cluster with the nearest mean.

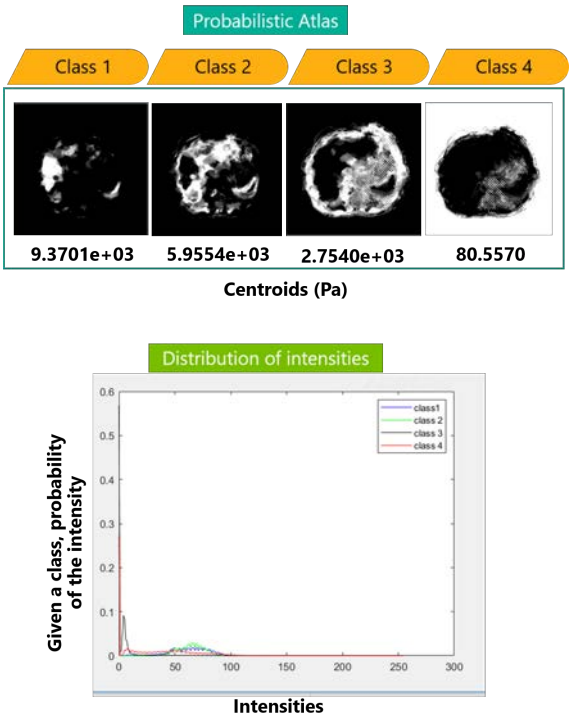


Figure 11: Probabilistic Atlas approach.

The centroids of the clusters (in Pascals) from the lowest to highest level of stiffness were 80,5570, 2.7540e+03, 5.9554e+03 and 9,3701 (see Figure 11) for patient 3. For any given pixel, a brighter intensity in the probabilistic atlas indicates a higher probability of belonging to a cluster. For example, it can be seen in the figure that class 4 is mostly related to the pixels in the background, while class 1 has a stronger presence in the right lobe of the liver. On the other hand, the distribution of intensities of each class is presented in the lower section of the same figure by a plot. Basically, the four classes have a flat distribution, except for a huge peak in the intensity zero, belonging to the distribution of class 4 (denoted by color red in the plot). This result means that given a class, the probability of having one intensity or another is almost the same.

In Figure 12, it is compared one axial slice of the elastography and its corresponding anatomical slice, slice 44, with respect to the stiffness estimation given by the atlas approach to the slice 1 of the same anatomical image, for which this property was not available. Basically, the estimated stiffness values in the slice 1 only reflect the prior probability estimation of the atlas, resulting in a segmentation almost identical to the one of slice 44, even though they are completely different images.

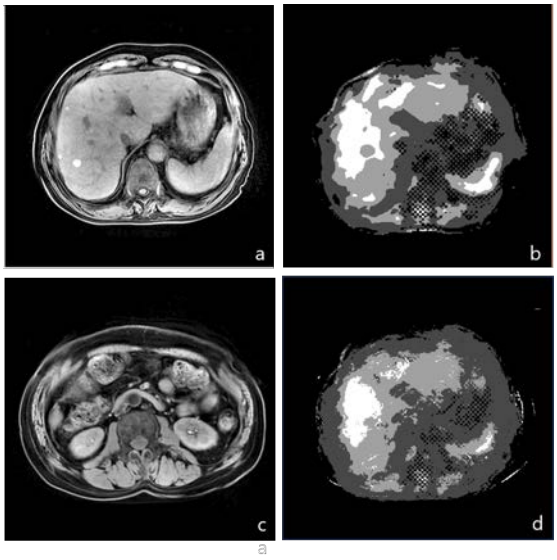


Figure 12: Images of patient 3: (a) Slice 44 of the volume in which the stiffness values are available and (b) its corresponding elastography slice, (c) Slice 1 of the volume in which the stiffness values are not available and (d) stiffness values assigned to it by the Probabilistic Atlas approach.

To end the sections of results, one slice of the volume of patient 3, in which the stiffness values are given by the elastography is exhibit, after the mask of the liver was used to isolate the hepatic tissue. The image shows the four clusters of stiffness. One of the clusters was almost eliminated by the mask, but it is still present in some of the pixels of the liver. Later on, the Z resolution of the volume of the liver is changed to match the high speed Mode of the printer of Stratasys, 0.027 mm between layers, and saved as a PNG bitmap using the simple Algorithm 3, making it ready for the printer.

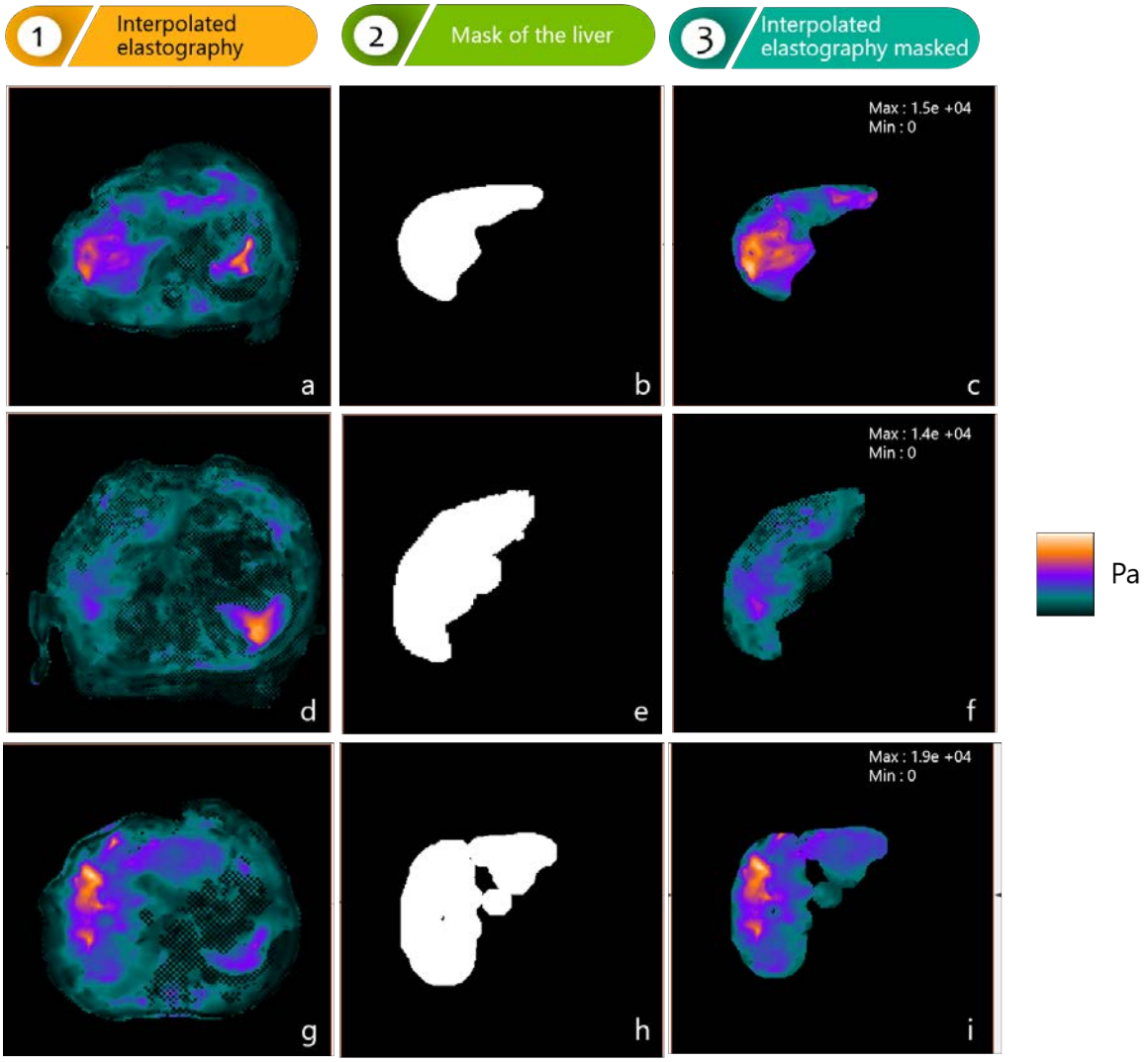


Figure 13: Images of patient 1: a, b, c; Images of patient 2: d, e, f; Images of patient 3: g, h, i.

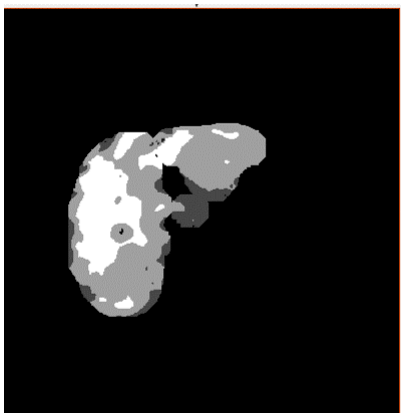


Figure 14: Volume of clustered stiffness is masked.

5. Discussion

In this section, the previous results are going to be interpreted based on what is already known and understand about the subject of this work. The pre-processing step provided the expected results, in terms of bias field correction and noise removal, facilitating the subsequent task of segmentation. The process of segmentation was time consuming because of the heterogeneous datasets of images, but the biggest difficulty was caused by the resolution of the volumes. MRI slices of 10 mm of thickness, for example cause a a non gradual transition between the slices. So, the different algorithms or even segmentation tools struggle to identify the tissues and perform the task of segmentation. Besides, if the intention is to take advantage of the many benefits of voxel-based printing, the quality of the images should be a priority.

Despite being challenging, the outcome of this step was

satisfactory and allowed to continue with the other steps of the framework. In general, one constraint of this thesis was the access to pairs of MRI and MRE images, since the second modality is still not widely used by hospitals.

After segmentation, assigning stiffness values of the liver was performed by using interpolation to make the elastography match the dimensions and resolution of the anatomical image and then masking the result with the mask of the liver. Stiffness of the liver presented significant variability depending on the region of measurement, specially in the case of patient 1 and patient 2. The mean liver stiffness was 6.39 kPa for patient 1, 3.71 kPa for patient 2 and 6.75 kPa for patient 3. The mean liver stiffness is used by physicians as an indicator of liver fibrosis, if the stiffness is superior to 3kPa, usually that is a sign of this disease in the patients (Venkatesh et al., 2013).

But the elastographies were prescribed so the liver was imaged in the widest portion, not the whole organ. For the remaining part, the atlas approach intended to estimate the stiffness properties. However, the results were not as desired. The probabilistic atlas and the distribution of intensities are combined using a Bayesian framework, to estimate the stiffness. The probability atlas corresponds to the prior probability, that is the probability of the hypothetical class 1, 2, 3 or 4 (of stiffness values) to occur, before the intensity is observed. The distribution represents the likelihood of an intensity (or the probability of observing an intensity, given a class). It indicates the compatibility of having an intensity with the given class. As appreciated in the result in Figure 11, since there is not a a proven relationship between intensity and stiffness, the likelihood did not provide enough evidence given the hypothesis of belonging to a any of the four classes. The final results only reflect the prior probability of belonging to a class, having an undesired performance. As a future work it would be interesting to investigate other approaches to estimate the stiffness.

For the remaining steps, the results of the probabilistic atlas were not considered. So, only one middle section of liver was re-slice and prepare for the final printing.

6. Conclusions

3D printing technologies are used to translate medical images into personalized physical models, revolutionizing the way in which physicians and scientists interact with medical data. Despite its great contribution to medical practice, the homogeneity of materials and colors offered by traditional 3D printing techniques simply fall short when trying to accurately replicate the enormous variety and complexity of tissues and organs.

In clear contrasts, the voxel-based framework described here provides a method for capturing complex properties of medical images in five stages: (1) Preprocessing: in which bias field correction and PCA are applied to reduce the noise in DICOM images. (2) Organ segmentation, the organ is segmented using global binary thresholding, morphological operations and the properties of the volumes. (3) Mapping of stiffness, the segmented organ is mapped to its equivalent stiffness value in the stiffness map 'elastography' using interpolation. (4) Re-slicing to printer Z resolution, the reconstructed 3D organ is prepared by re-slicing it by parts to match the printer resolution. (5) Bitmap generation, to print the organ voxel by voxel with their respective stiffness value by choosing a bitmap as the input data type.

We hope that this methodological effort contributes to accelerate the rate of adoption of this new type of technology by a wide range of medical and scientific professionals who require more precise anatomical models.

7. Acknowledgments

I would like to express my deepest appreciation to my supervisors Dr. Oliver Díaz and Dr. Robert Martí for their continues guidance, support and motivation during the master thesis project.

References

Agarwala, S., 2016. A perspective on 3d bioprinting technology: present and future. American Journal of Engineering and Applied Sciences 9, 985–990.

Bader, C., Kolb, D., Weaver, J.C., Oxman, N., 2016. Data-driven material modeling with functional advection for 3d printing of materially heterogeneous objects. 3D Printing and Additive Manufacturing 3, 71–79.

Bader, C., Kolb, D., Weaver, J.C., Sharma, S., Hosny, A., Costa, J., Oxman, N., 2018. Making data matter: Voxel printing for the digital fabrication of data across scales and domains. Science advances 4, eaas8652.

Brunton, A., Arikan, C.A., Urban, P., 2015. Pushing the limits of 3d color printing: Error diffusion with translucent materials. ACM Transactions on Graphics (TOG) 35, 4.

Chae, M.P., Rozen, W.M., McMenamin, P.G., Findlay, M.W., Spychal, R.T., Hunter-Smith, D.J., 2015. Emerging applications of bedside 3d printing in plastic surgery. Frontiers in surgery 2, 25.

Chang, D., Tummala, S., Sotero, D., Tong, E., Mustafa, L., Mustafa, M., Browne, W.F., Winokur, R.S., 2019. Three-dimensional printing for procedure rehearsal/simulation/planning in interventional radiology. Techniques in vascular and interventional radiology 22, 14–20.

Clark, M., Ghammraoui, B., Badal, A., 2016. Reproducing 2d breast mammography images with 3d printed phantoms, in: Medical Imaging 2016: Physics of Medical Imaging, International Society for Optics and Photonics. p. 97830B.

Community, G., 2019. How to make grabcad voxel print slices using matlab. `https://grabcad.com/tutorials/how-to-make-grabcad-voxel-print-slices-using-matlab`.

Doubrovski, E.L., Tsai, E.Y., Dikovsky, D., Geraedts, J.M., Herr, H., Oxman, N., 2015. Voxel-based fabrication through material property mapping: A design method for bitmap printing. Computer-Aided Design 60, 3–13.

Farooqi, K.M., Saeed, O., Zaidi, A., Sanz, J., Nielsen, J.C., Hsu, D.T., Jorde, U.P., 2016. 3d printing to guide ventricular assist device placement in adults with congenital heart disease and heart failure. JACC: Heart Failure 4, 301–311.

Floyd, R.W., 1976. An adaptive algorithm for spatial gray-scale, in: Proc. Soc. Inf. Disp., pp. 75–77.

GE, 2018. Lava flex. http://www3.gehealthcare.com.sg/en-gb/products/categories/magnetic_resonance_imaging/body_imaging/lava_flex.

Gonzalez, R.C., Woods, R.E., et al., 2002. Digital image processing [m]. Publishing house of electronics industry 141.

Gubern-Mérida, A., Kallenberg, M., Martí, R., Karssemeijer, N., 2011. Multi-class probabilistic atlas-based segmentation method in breast mri, in: Iberian Conference on Pattern Recognition and Image Analysis, Springer. pp. 660–667.

HarvardMedicalSchool, 2019. About slicer. https://www.slicer.org/.

Hawley, J.R., Kalra, P., Mo, X., Raterman, B., Yee, L.D., Kolipaka, A., 2017. Quantification of breast stiffness using mr elastography at 3 tesla with a soft sternal driver: A reproducibility study. Journal of Magnetic Resonance Imaging 45, 1379–1384.

Hosny, A., Keating, S.J., Dilley, J.D., Ripley, B., Kelil, T., Pieper, S., Kolb, D., Bader, C., Pobloth, A.M., Griffin, M., et al., 2018. From improved diagnostics to presurgical planning: high-resolution functionally graded multimaterial 3d printing of biomedical tomographic data sets. 3D Printing and Additive Manufacturing 5, 103–113.

Juntu, J., Sijbers, J., Van Dyck, D., Gielen, J., 2005. Bias field correction for mri images, in: Computer Recognition Systems. Springer, pp. 543–551.

Karthikeyan, A., Valliammai, M., 2012. Lungs segmentation using multi-level thresholding in ct images. Int. J. Electron. Comput. Sci. Eng 1, 1509–1513.

Low, G., Kruse, S.A., Lomas, D.J., 2016. General review of magnetic resonance elastography. World journal of radiology 8, 59.

Maragiannis, D., Jackson, M.S., Igo, S.R., Schutt, R.C., Connell, P., Grande-Allen, J., Barker, C.M., Chang, S.M., Reardon, M.J., Zoghbi, W.A., et al., 2015. Replicating patient-specific severe aortic valve stenosis with functional 3d modeling. Circulation: Cardiovascular Imaging 8, e003626.

Mariappan, Y.K., Glaser, K.J., Ehman, R.L., 2010. Magnetic resonance elastography: a review. Clinical anatomy 23, 497–511.

MathWorks, T., 2019. Matlab app designer. https://www.mathworks.com/products/matlab/app-designer.html.

Medixant, 2019. Radiant dicom viewer. https://www.radiantviewer.com/.

Murphy, M.C., Huston, J., Glaser, K.J., Manduca, A., Meyer, F.B., Lanzino, G., Morris, J.M., Felmlee, J.P., Ehman, R.L., 2013. Preoperative assessment of meningioma stiffness using magnetic resonance elastography. Journal of neurosurgery 118, 643–648.

Ploch, C.C., Mansi, C.S., Jayamohan, J., Kuhl, E., 2016. Using 3d printing to create personalized brain models for neurosurgical training and preoperative planning. World neurosurgery 90, 668–674.

Slicer, D., 2013. N4itk mri bias correction. https://www.slicer.org/wiki/Documentation/4.3/Modules/N4ITKBiasFieldCorrection.

Solomon, J., Ba, A., Bochud, F., Samei, E., 2016. Comparison of low-contrast detectability between two ct reconstruction algorithms using voxel-based 3d printed textured phantoms. Medical physics 43, 6497–6506.

Stanford, 2019. Imaging clinic at stanford medicine imaging center in palo alto. https://stanfordhealthcare.org/medical-clinics/imaging-clinic-stanford-medicine-imaging-center.html.

Stratasys, 2017. Lancaster university pushes design limits with grabcad voxel print technology. https://www.stratasys.com/resources/search/case-studies/lancaster-university?resources=cd10362d-b310-4078-a6ea-82a25813ef06&phrase=lancaster.

Stratasys, 2019. J735 and j750. https://www.stratasys.com/3d-printers/j735-j750.

Sullivan, J., Miller, R., Pios, G., 1993. Image halftoning using a visual model in error diffusion. JOSA A 10, 1714–1724.

Tang, A., Cloutier, G., Szeverenyi, N.M., Sirlin, C.B., 2015. Ultrasound elastography and mr elastography for assessing liver fibrosis: part 1, principles and techniques. American journal of roentgenology 205, 22–32.

Venkatesh, S.K., Wells, M.L., Miller, F.H., Jhaveri, K.S., Silva, A.C., Taouli, B., Ehman, R.L., 2018. Magnetic resonance elastography: beyond liver fibrosis—a case-based pictorial review. Abdominal Radiology 43, 1590–1611.

Venkatesh, S.K., Yin, M., Ehman, R.L., 2013. Magnetic resonance elastography of liver: technique, analysis, and clinical applications. Journal of magnetic resonance imaging 37, 544–555.

Weeger, O., Kang, Y.S.B., Yeung, S.K., Dunn, M.L., 2016. Optimal design and manufacture of active rod structures with spatially variable materials. 3D Printing and Additive Manufacturing 3, 204–215.

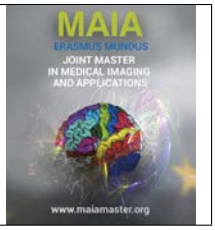
Zhang, Y., Ge, H.w., Li, N.c., Yu, C.f., Guo, H.f., Jin, S.h., Liu, J.s., Na, Y.q., 2016. Evaluation of three-dimensional printing for laparoscopic partial nephrectomy of renal tumors: a preliminary report. World journal of urology 34, 533–537.

Appendix I: Software Packages for 3D Reconstruction and Mesh Reparation

NAME	A) Company	B) Threshold/ Segmentation	C) Export STL	D) Repair STL	E) Free	F) OS Platform
Mimics	Materialise	Yes	Yes	No	No	Mac, Windows
3D Slicer	Surgical Planning Laboratory	Yes	Yes	No	Yes	Mac, Windows
Seg3D	University of Utah	Yes	Yes	No	Yes	Mac, Windows
MITK	German Cancer Research Centre	Yes	Yes	No	Yes	Mac, Windows
Osirix	Pixmeo	Yes	Yes	No	Yes	Mac
InVesalius	CTI Renato Archer	Yes	Yes	No	Yes	Mac, Windows
MeVisLab	MeVis Medical Solutions AG	Yes	Yes	No	Yes	Mac, Windows
MIPAV	NIH CIT	Yes	Yes	No	Yes	Mac, Windows
3D Doctor	Able Software	Yes	Yes	No	No	Windows
Dolphin 3D	Software Imaging and Management	Yes	Yes	No	No	Windows
ScanIP	Synopsys	Yes	Yes	No	No	Windows
Analyze	AnalyzeDirect	Yes	Yes	No	No	Mac, Windows
MeshLab	The Visual Computing Lab	No	Yes	Yes	Yes	Mac, Windows
MeshMixer	AutoDesk Research	No	Yes	Yes	Yes	Mac, Windows
3-Matic	Materialise	No	Yes	Yes	No	Mac, Windows
3D Builder	Microsoft	No	Yes	Yes	Yes	Windows
Rhinoceros	McNeel	No	Yes	Yes	No	Mac, Windows
FreeCAD	The FreeCAD team	No	Yes	Yes	Yes	Mac, Windows
LimitState:Fix	LimitState:FIX	No	Yes	Yes	No	Mac, Windows

Medical Imaging and Applications

Master Thesis, June 2019



Breast MRI Normalization to Predict Pathological Complete Response to Neoadjuvant Chemotherapy

Fahad Khalid, MAIA team

Erasmus Mundus Joint Master Degree in Medical Imaging and Applications
University of Girona, Spain; University of Cassino, Italy; University of Burgundy, France

Frederique Frouin

IMIV Laboratory, Inserm-CEA, Service Hospitalier Frederic Joliot, Orsay, France

Abstract

Purpose: The aim of the study is to standardize different MRI acquisitions in order to achieve robust comparisons between exams done on different subjects as well as between exams performed on different times in the same subject.

Methods: To achieve this task, we propose a robust image processing pipeline to extract a reference structure. For the present study, we used T1 weighted Dynamic Contrast Enhanced MRI and T2 weighted scans of 44 patients with breast tumor diagnosis, that were acquired at the Institut Curie, Paris, France. Firstly, we explored the N4ITK bias field correction method to improve signal homogeneity in T1 and T2 weighted images. Then we implemented a dedicated method to extract the subcutaneous fat layer from T1 images.

Results: We proposed some new default parameters for bias field correction using N4ITK, that improve the uniformity inside the whole field of view for both T1 and T2 images. The pipeline for subcutaneous fat extraction was successfully applied to the 44 T1 weighted scans. An histogram analysis inside the subcutaneous fat layers revealed two different patterns, which were due to the use of two different coils.

Conclusion: This study led us to conclude that MR bias field correction is an important factor to better quantify breast MR images. Some further investigation is necessary to recover some more comparable signal between the two different coils.

Keywords: Breast MR, Radiomics, MR bias field correction, Fat layer extraction

1. Introduction

Radiomics makes use of data characterization algorithms to extract meaningful qualitative features from medical images. It aims at establishing more developed and precise patient diagnosis, staging cancers, determining optimum therapies, predicting patient outcomes or their risk level, or choosing the radiation therapy dose level as described in Lambin et al. (2012), Kumar et al. (2012), Gillies et al. (2015), Parekh and Jacobs (2016). With the increase of the number of breast cancer cases, studies suggest that a more precise and characterized approach would be beneficial for advancement in breast cancer therapy. Radiomics helps in revealing tumor characteristics or predicting prognosis through the extraction of a great number of imaging indices inside the tumor area (Lambin et al. (2012), Gevaert et al. (2014), Aerts (2016)). Breast cancer is a tumour which develops from the cells that form the mammary gland. Breast cancer (BC) was the leading cancer location in women in all European countries in 2012, and also the main cause of death from cancer in women in Europe. In

2017, about 250,000 new cases of invasive breast cancer were diagnosed, and 60,000 cases of in situ breast carcinoma. Approximately 40,000 women died from breast cancer in 2017 (Win). Prevalence of BC is increasing due to early diagnosis and changes in risk factors, but also to aging of the population. It is a heterogeneous disease, with different molecular sub-types. Each type of breast tumor calls for a specific treatment. Even with treatment, most patients with locally advanced breast cancer will develop (Giordano (2003)) in their work point out, even after treatment most tumors are likely to develop distant metastases.

Neoadjuvant chemotherapy (NAC) is often a first line of defense in the treatment of locally advanced breast cancer. Proposed to patients prior to surgery, NAC is known to reduce tumor extent, improve patients surgical outcomes, and shrink metastasis grow (Thompson and Moulder-Thompson (2012)). The ideal outcome of NAC is the pathologic complete absence of residual invasive tumor cells within excised breast tissue following NAC, or pathological complete response (pCR), which strongly predicts favorable prognosis as compared with patients who experience partial or no response (non-pCR) (Luangdilok et al. (2014), Kong et al. (2011)). Less than 10% to 50% of breast cancer patients undergoing NAC achieve pCR, and thus there is a need for reliable noninvasive pre-treatment predictors of pCR that can enable better and smarter procedures of NAC and prevent a delay in effective treatment for patients with non-responding or progressing tumors.

Breast magnetic resonance (MR) imaging is often acquired for patients near diagnosis of breast cancer as it is an important component in clinical work-up (Knuttel et al. (2014), Brasic et al. (2013)). Dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) offers visual differentiation of lesions from normal tissue owing to the increased vascularisation and capillary permeability of breast lesions (Shin et al. (2011), Belli et al. (2006)). Therefore dynamic MR imaging is more likely a modality that is possibly complementary to mammography and ultrasonography (US) because of the additional three-dimensional spatial and temporal information about the lesion that it yields. Because of its high sensitivity to tumor presence and angiogenic changes, dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) is the preferred imaging modality in the NAC setting and has been demonstrated to effectively predict pCR following an early treatment period. For instance, changes in volumetric and kinetic parameters perform well in pCR prediction but afford no intuition with regard to pCR prior to treatment (Marinovich et al. (2012), Delille et al. (2003), Padhani et al. (2006), Dialani et al. (2015)). There remains a shortage of reliable clinical pCR indicators based on DCE-MRI that do not require previous NAC administration (Marinovich et al. (2012), Dialani et al. (2015)). An increasing advancement in technology has given rise to several MR system manufacturers, and the various systems have different acquisition systems, magnetic field strengths, pulse sequences, and coils that continue to be modified and improved over time. This brings a large number of challenges for image analysis algorithms and associated radiomics studies. Many sources of variability can affect the results, especially in a large multi-center clinical study. Indeed, standardization of breast MRI signal intensity is not widespread, even if the acquisition protocols are quite normalized. There is no universally applied quality assurance procedures to ensure robust standardization of MR scans coming for different scanners. The present study aims at proposing some new procedures to better standardize images issued from clinical studies.

2. State of the art

Keeping in mind the challenges posed by MR image noise and the presence of a slow-varying background component of MR image non-homogeneity, and non-standardization of MRI, the present section will review the currently literature available to address these issues. Sufficient literature is available documenting successive completion of brain MR standardization and homogeneity algorithms. Breast MR analysis on the other hand is still in its initial stage and poses interesting challenges for researchers due to the vast variation in the breast physiology, including for instance breast tissue density, hormonal status, and age. Furthermore, the non-homogeneity inherent to the radio-frequency coils is usually translated into the reconstructed MR images. This non-homogeneity is described as the bias field signal. Image processing algorithms such as segmentation, texture analysis or classification that use the gray level values of image pixels will not produce satisfactory results if the images are not bias field corrected. A preprocessing step is thus needed to correct for the bias field signal before submitting corrupted MRI images to such algorithms. Many works have proposed bias field correction methods that are integrated into tissue classification algorithms, typically within the domain of brain MRI analysis (Wells et al. (1996), Held et al. (1997) Van Leemput et al. (1999), Zhang et al. (2001)). In their work, Ahmed et al. (2002) successfully modify the fuzzy C-means algorithm to achieve bias field correction. Then came the N3 bias field correction proposed by Larsen et al. (2014). The N3 method is iterative and seeks the smooth multiplicative field that maximizes the high frequency content of the distribution of tissue intensity. The method is fully automatic, requires no *a priori* knowledge and can be applied to almost any MR image. More recently, Tustison et al. (2010) proposed an improved version of the N3 algorithm known as the *N4ITK*. The improvement from N3 to N4 was achieved by replacing the B-spline smoothing strategy used in the original N3 framework with an advantageous alternative

(Lee et al. (1997), Tustison and Gee (2005), Tustison and Gee (2006)), which addresses major issues explored by previous N3 evaluation studies. As mentioned previously in this section the bias correction methods have been applied to brain MR image data-sets. Not much documentation is published for experiments exploring the breast MR data-sets. In our work we explored the N4 bias field correction for breast data set under study, in order to improve results when compared to a conventional setting of parameters in the *N4ITK* algorithm. A second aim of our work was to achieve segmentation of a reference tissue of the breast to further standardize image intensities using histogram analysis of gray levels inside this reference tissue. Hence we review attempts to achieve brain MR intensity standardization by image histogram analysis. Interesting and extensive reviews were given in (Madabhushi et al. (2006)), (Jager and Hornegger (2008)), (Shah et al. (2011)). The approaches used by (Christensen (2003)), implies using even-ordered derivatives of the image histogram which results in a single global scaling factor between the two images. Weisenfeld and Warfield (2004) used Kullback-Leibler divergence to match the intensity distribution of two images. Leung et al. (2010) proposed a semi automated segmentation technique to delineate the three main brain tissue components (grey matter, white matter, cerebrospinal fluid) followed by computing mean intensities to realign the whole intensities in the images. The drawback of this method is that it yields a linear transformation, which does not completely address the problem, guaranteeing the standardization of spatially corresponding tissue intensities (Robitaille et al. (2012)). In the study of (Jager and Hornegger (2008)), the properties of all acquired images (e.g., T1- and T2-weighted images) are stored in multidimensional joint histograms. In order to normalize the probability density function of a newly acquired dataset, a nonrigid image matching is performed between the joint histogram of a reference and the joint histograms of the newly acquired images, avoiding any prior registration or segmentation of the datasets (Jager and Hornegger (2008)).

Figure 1: Breast MRI acquisition (reprint from https://www.mayoclinic.org/tests-procedures/breast-mri/about/pac-20384809)

3. Material and methods

3.1. *MR Image Data:*

The image data set consists of 44 axial T2 weighted images and T1 weighted DCE-MRI scans of patients diagnosed with breast tumor, at Institute Curie (Paris, France). All breast MR imaging examinations were performed within one week before initiating NAC. For T2 imaging, Dixon sequences were acquired and fat-suppressed images were further analyzed. For T1 imaging, DCE sequences were recorded after an initial fat-saturated T1-weighted pre-contrast scan. After an intravenous injection of 0.2 ml/kg gadolinium contrast agent, the first post-contrast scan was collected within 2 minutes. The acquisition is performed as the patient lies on her stomach as show in Fig. [1]. All the scans were acquired using a *Siemens* 1.5 T MR scanner with two different coils: a coil dedicated to breast imaging, *constructor breast coil*, and a coil dedicated to breast biopsy *Sentinelle biopsy coil*. Most of the scans (77%) were acquired using the biopsy coil, while the remaining scans (23%) were acquired using the constructor coil. Both coils are double.

3.2. ***Normalization:*** *N4 Bias field correction*

Breast MR acquisition is quite unique due to its double coil nature. A definitive confounding factor in MR acquisition is the disturbance of the low frequency non uniformity present in the image data know as bias (see Fig. [2] for an illustration). The N4 bias field correction algorithm has proven to be quite effective for brain MR scans, but the default hyper parameters were not suited for our data. When testing these default values, there was a change in intensity values but the bias field was quasi-uniform in the whole field of view. The image formation model used by *N4ITK*, N3 and other bias field correction algorithms is:

$$v(x) = u(x)f(x) + n(x), \tag{1}$$

where x is the voxel, v is the input MR volume, u is the output bias corrected image volume, f is the estimated bias field, and n is the noise (considered as Gaussian and independent).

The logarithmic transformation (notation $\hat{u} = \log u$) is frequently used to work with the noise-free volumes:

$$\hat{v}(x) = \hat{u}(x) + \hat{f}(x). \tag{2}$$

The bias corrected volume is obtained iteratively. At the n^{th} iteration:

$$\begin{aligned} \hat{u}^n &= \hat{v} - \hat{f}_e^n, \\ &= \hat{v} - \mathbf{S}\left\{\hat{v} - \mathbf{E}[\hat{u}|\hat{u}^{n-1}]\right\}, \end{aligned} \tag{3}$$

where $\hat{u}^0 = \hat{v}$, $\hat{f}_e^0$ (the initial bias field estimate) is set to 0 and $\mathbf{S}$ is the smooth B-spline approximation, which

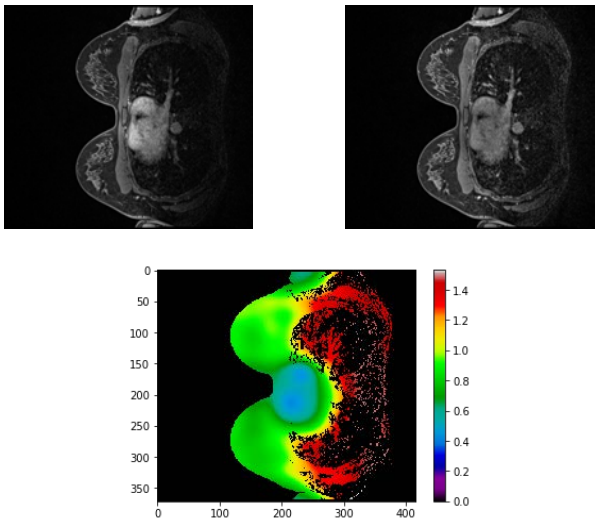


Figure 2: Top left: T1 DCE MR image, showing some bias field. Top right: T1 DCE MR image after the application of the bias field correction algorithm, showing more homogeneity, especially in the regions of the back. Bottom: estimation of the bias field.

was proposed by the N3 algorithm. The N4 algorithm uses an improved version of the iterative estimator, that is expressed by equation (4):

$$\begin{aligned} \hat{u}^n &= \hat{u}^{n-1} - \hat{f}_r^n \\ &= \hat{u}^{n-1} - \mathbf{S}^*\left\{\hat{u}^{n-1} - \mathbf{E}[\hat{u}|\hat{u}^{n-1}]\right\}. \end{aligned} \tag{4}$$

The $\mathbf{S}^*$ is an improved B-spline approximation and $\hat{f}_e^n$ is an estimation of the residual bias field at the n^{th} iteration.

We conducted several experiments by varying the hyper parameters such as the *number of iterations* and *number of fitting levels*. Their default values (optimized for brain studies) were equal to 50 and 4. For our breast database, the best parameters while keeping the computation time acceptable were found to be:

- Number of iterations = 50
- Number of fitting levels = 5
- Use of a mask to reduce the computation of the bias field inside this mask.

To ensure the robustness of *N4ITK* with the hyper parameters we tested the algorithm on the different scans available in our data set.

3.3. Mean contrast analysis

To validate the performance of the N4 bias correction, we performed a statistical analysis on the whole database. We selected four regions in each MR volume: two locations in the right breast (r) and two locations in the left breast (l) were delineated by an expert radiologist. Two regions in pectoral muscles (P_r and P_l) and two regions in the normal breast parenchyma (NB_r and NB_l) were identified and marked (Fig.[3]). The regions were delineated using the *LIFEx* software.

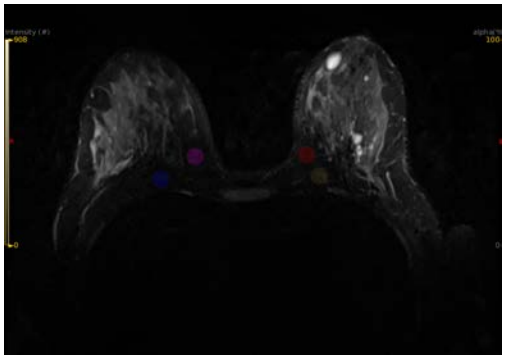


Figure 3: Examples of the regions of interest that are used for the calculation of contrast: blue color is for the right pectoral muscle, pink color for the right normal breast, red color for the left normal breast region, and yellow for the left pectoral muscle

Then the contrast between the left and right pectoral muscle (C_P), and the contrast between the left and right parenchyma (C_{NB}) were computed for all the MR volumes before and after applying *N4ITK* bias correction, as shown in (5):

$$\begin{aligned} C_P &= 2|P_r - P_l|/(P_r + P_l) \\ C_{NB} &= 2|NB_r - NB_l|/(NB_r + NB_l) \end{aligned} \tag{5}$$

3.4. Pipeline for breast subcutaneous fat layer extraction

T1-weighted fat-saturated MRI images are usually acquired in clinical breast MRI imaging protocols and are used for breast segmentation and density estimation. In T1 weighted DCE images, fat appears as the brightest along with certain vessels and and tumor region. The tumor and vessels can be identified as very bright locations. They are rarely in the subcutaneous fat layer. Fibro-glandular tissue and the chest wall appear as moderate signals with quite similar signal intensity. Hence, for challenging cases where a part of the fibro-glandular tissue is connected to the chest wall and when there is no fat along the anterior side of the chest wall, breast segmentation is a difficult task. There are a few studies which suggest automated approaches for breast segmentation. Ertaş et al. (2008) presented a breast segmentation method using two different cellular neural networks: the first one is to perform some threshold operations, and the second one is for removing small objects and smoothing sharp corners. Ertas et al. (2017) presents 3D bias-corrected fuzzy c-means clustering and morphological operation to extract the breast region. Hayton et al. (1997) proposed iterative application of morphological opening operator, with an increasing scale at each iteration until eliminating the breast region and keeping the initial approximation of

breast wall. In the development of our pipeline, a special attention was given on making sure that the method is robust and segments the subcutaneous layer only. The well known image processing libraries *skimage* and *scipy* were used for most of the image analysis operations along with a few in-house developed functions. As the MR images were in *nifti* format, the python library *nibable* was used to handle the nifti format.
The subcutaneous fat layer is the most outer layer of the internal breast, it is located just beyond the skin. Hence to extract it, we propose to isolate the contour of the breast as shown in Fig. [4] .

- **Non-homogeneity Correction**;
  The original volumes in our data set showed significant impact of the bias present in them. As stated earlier from the literature review we know that introducing homogeneity was an important step. We achieved this by the optimized hyper parameters. To work with more comparable intensity values, each volume was then normalized to set its maximal value equal to 1.

- **Mask Generation**;
  We used a threshold (the threshold value was the mean intensity of the volume) to generate the mask of the whole breast region. This threshold was defined in order to isolate the chest from the background. This is a rough step, since the MR images show low intensity values in the lung for instance, and the idea of this first step is to have a rough contour of the chest. Hence due to low values in lung for instance, the binary mask (i.e. composed of binary values: 0 and 1, an example of which can be found in Fig [4]) obtained from the threshold application contains some holes. To remove some of them without degrading the breast contour, we use a morphological operation of hole filling. The algorithm used in this function consists in invading the complementary of the shapes in input from the outer boundary of the image, using binary morphological dilations with a circular structuring element size 1. Holes which are not connected to the boundary are not invaded. The result is the complementary subset of the invaded region.

- **Extraction of subcutaneous fat layer** ;
  Once we have obtained a complete mask, f, of the breast region the next step is to distinguish between the values lying at the contour of the breast from the center of the chest. Morphological Euclidean distance transformation is then used. The distance transformation provides a measure of the distance of each voxel inside the binary mask to the contour of the mask.
  Distance transform algorithm:

  - To compute the distance of each voxel (i, j, k) to the background S :
  - At iteration n, compute $F_n[i, j, k]$:
  - $F_0[i, j, k] = f[i, j, k]$ (initial values)
  - $F_n[i, j, k] = F_0[i, j, k] + min(F_{n-1}[u, v, w])$, (u, v, w) being the 6-neighbor voxels of (i, j, k) that is voxels such as $D([i, j, k], [u, v, w]) = 1$
  - Repeat iterations ($n = n + 1$) until no change is observed between F_n and F_{n-1}

  Finally a new binary mask, g, was defined from the previous mask f, by keeping the voxels being at a distance less than 2 mm from the outer contour of the mask. The idea is that only these voxels of the mask g could be considered as possible subcutaneous fat voxels.

- **Final segmentation**;
  The binary image g contains the subcutaneous fat layer but also some unwanted objects such as lung, heart, and in some cases parts of the arms. We multiply the mask g with the original volume (after bias field correction) in order to retrieve the voxel intensity values. Then we used a hysteresis threshold, the low threshold being equal to 0.2 and the high threshold being equal to 0.95 (for a maximum being equal to 1). From practical experimentation, the resulting voxels should correspond to subcutaneous fat layer and nothing else. Finally a binary mask h is generated with only the subcutaneous layer as 1's and everything else set to 0.

3.5. Histogram Analysis for T1 DCE standardization

Image histograms display a graphical representation of the gray level or tonal distribution in a digital image. It plots the number of pixels for each interval of intensity values (bin). For MR breast radiomics, we chose to study the subcutaneous fat layer on T1 DCE images, in order to further standardize image intensities, similarly to what has been proposed in the IMIV laboratory for brain cancer MR images (Goya-Outi et al. (2018)). The three main reasons for choosing the subcutaneous fat layer for reference tissue are the following:

1. This region should have high signal intensities.
2. Regions containing tumor should be avoided as every tumor is unique and may disrupt the standardization process.
3. The subcutaneous fat layer seems better than the normal fat inside the breast region, which may be a mixture of different tissues.

The above factors are taken into consideration so as to make sure that the reference tissue we chose to extract and standardize is robust enough and comparable in all MR breast scans. After the extraction of the subcutaneous fat layer, we compute histograms inside it and consider that the first peak is a good candidate to represent subcutaneous fat tissue.

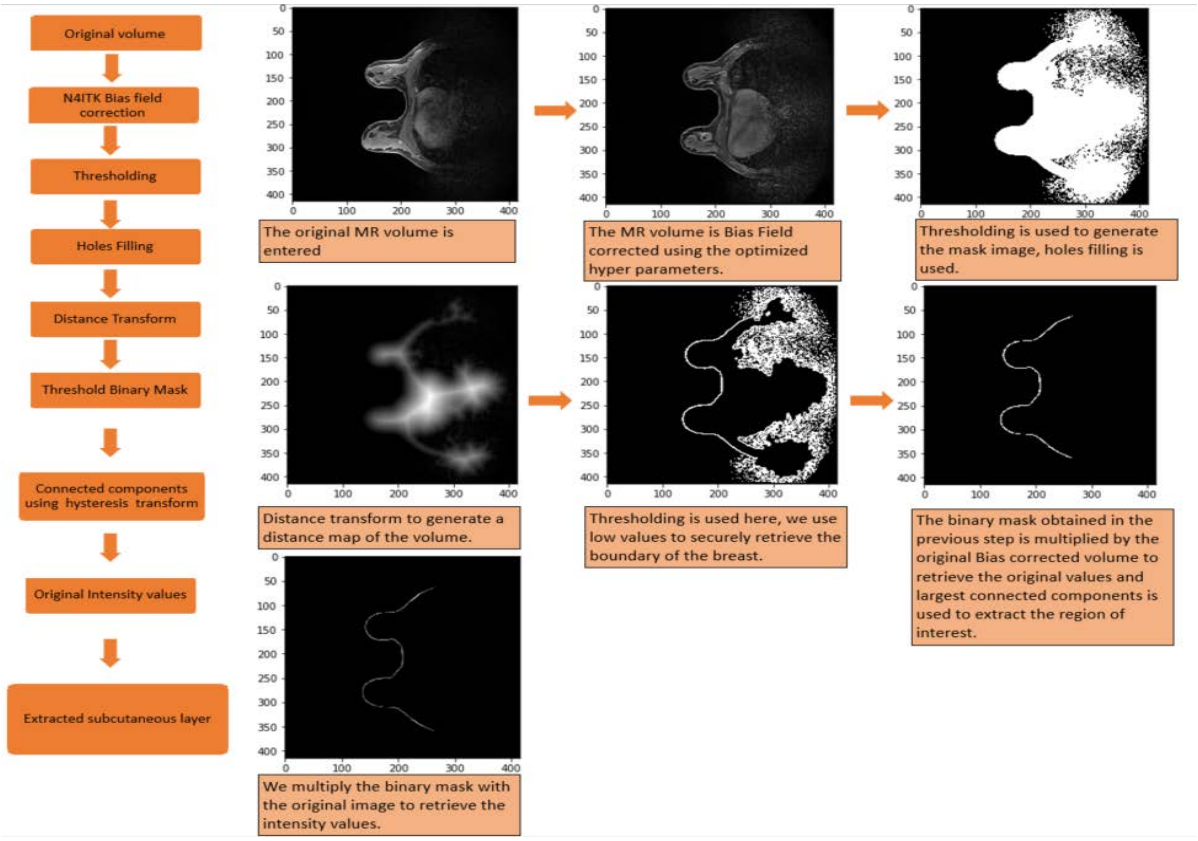


Figure 4: Pipeline for subcutaneous layer extraction

4. Results

4.1. Illustration of N4 Bias field correction

The results are presented in the next figures. Applying default parameters for N4 algorithm did not produce satisfactory results. Furthermore the N4ITK uses Otsu's method to build the mask image inside which the bias field is computed. It happens that this mask does not cover the entire breast area. Fig. [5] displays an example of the application of the N4ITK on T1 DCE MRI with default parameters. When comparing the original image with the N4 bias field corrected image, it appears that a very minor bias is estimated, and that better results could be expected. The estimation of the bias field shows very low levels of non-homogeneity. Results obtained using optimized hyper parameters for N4ITK are illustrated for three groups of scans: 1) T1 DCE MR scans acquired using the Sentinelle coil (see Fig. [6]), 2) T2 MR scans acquired using the Sentinelle coil (see Fig. [7]), and 3) T1 DCE MR scans acquired using the constructor coil (see Fig. [8]). The algorithm has proved to be efficient for all these three cases. There is a significant modification in intensity values after the implementation of the algorithm and a non uniform bias field was estimated for each case.

4.2. Experimental validation of N4 Bias field correction

Ideally, the contrast between the left and right sides should be close to zero. A global reduction of the mean values of contrast was observed after N4 correction in both T1 and T2 images. To display it, Fig. [9] and Fig. [10] represent box plots of the contrast values before and after N4, for T1 and T2 images, in the pectoral muscle and in the normal breast tissue. All the contrast values were reduced after N4 correction: indeed, for T1 DCE scans, mean contrast values in the pectoral muscle were equal to 0.122 before N4 and reduced to 0.085 after N4 correction. In the normal breast tissues, contrast was equal to 0.228 before N4 and to 0.107 after N4 correction. For T2 scans, mean contrast values in the pectoral muscle were equal to 0.156 before N4 and to 0.121 after N4 correction. In the normal breast tissues, contrast was equal to 0.219 before N4 and to 0.127 after N4 correction.

4.3. Extraction of subcutaneous fat layer

The pipeline described in the previous section has been tested and evaluated using the 44 T1 DCE volumes (see Fig. [11] for some illustrations). The segmentation is quite satisfactory as it shows robustness across our data set. The extracted subcutaneous fat layer was then submitted to histogram analysis. From the literature review of brain MR analysis, we hoped to observe

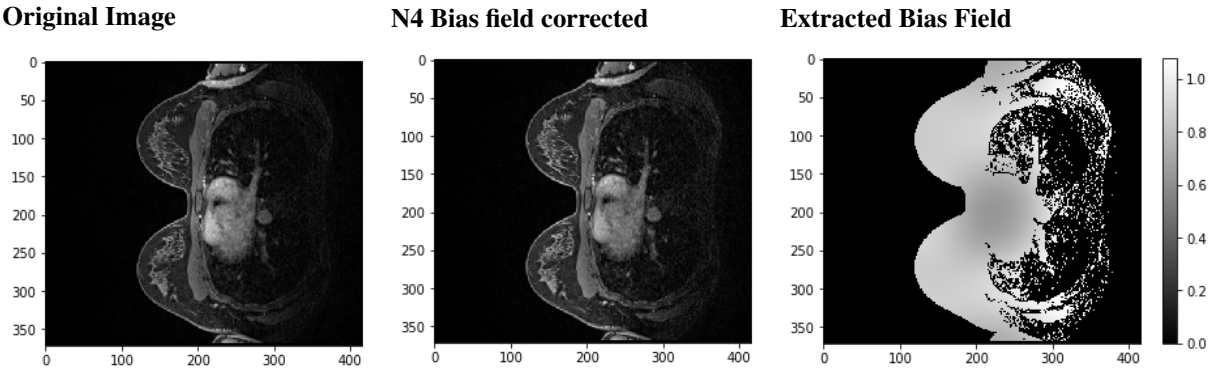


Figure 5: T1 DCE MRI after bias correction using default hyper parameters, scan acquired using *Sentinelle biopsy coil*

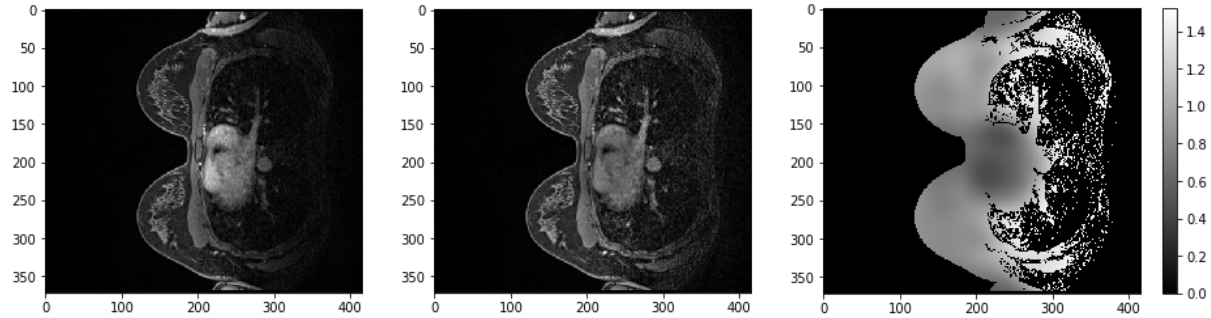


Figure 6: T1 DCE MRI after bias correction using optimized hyper parameters, scan acquired using *Sentinelle biopsy coil*

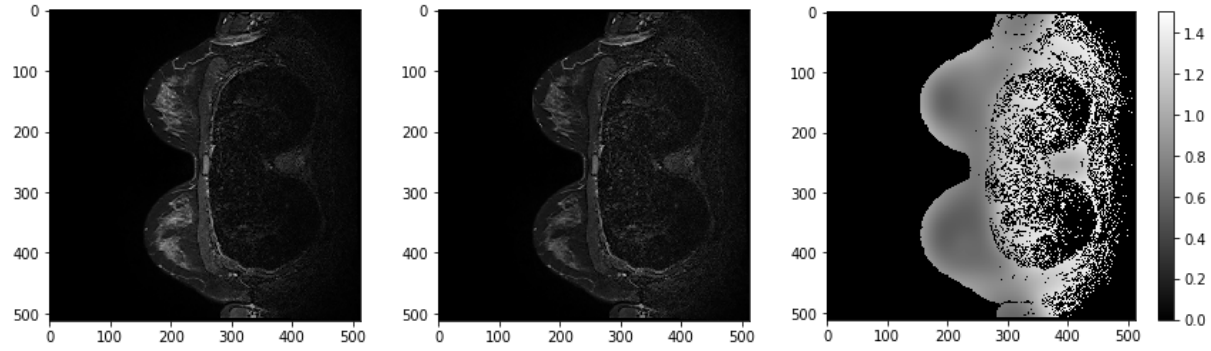


Figure 7: T2 weigthed MRI after bias correction using optimized hyper parameters, scan acquired using *Sentinelle biopsy coil*

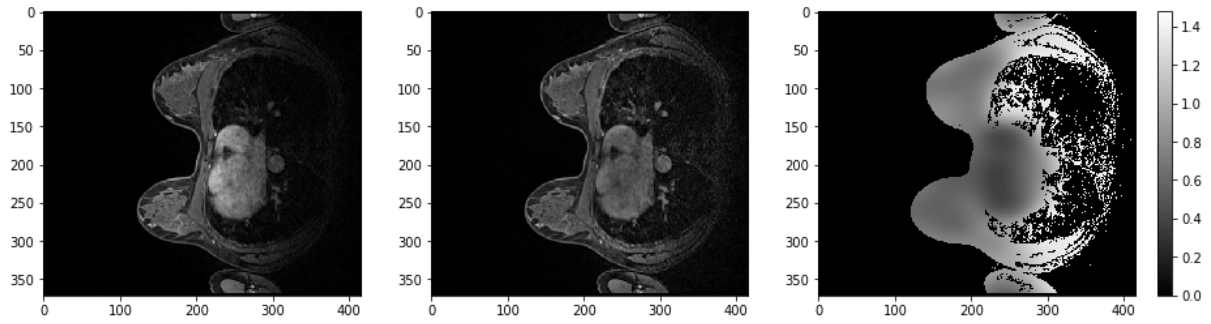


Figure 8: T1 DCE MRI Bias correction using optimized hyper parameters, scan acquired using the *constructor breast coil*

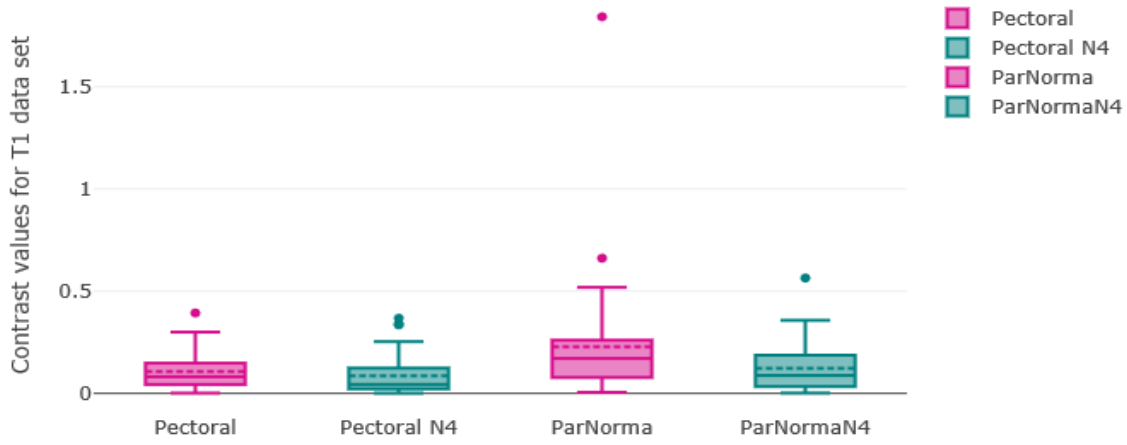


Figure 9: Left-Right contrast in the pectoral muscle and in the normal breast before and after N4ITK correction. Case of T1 weighted images

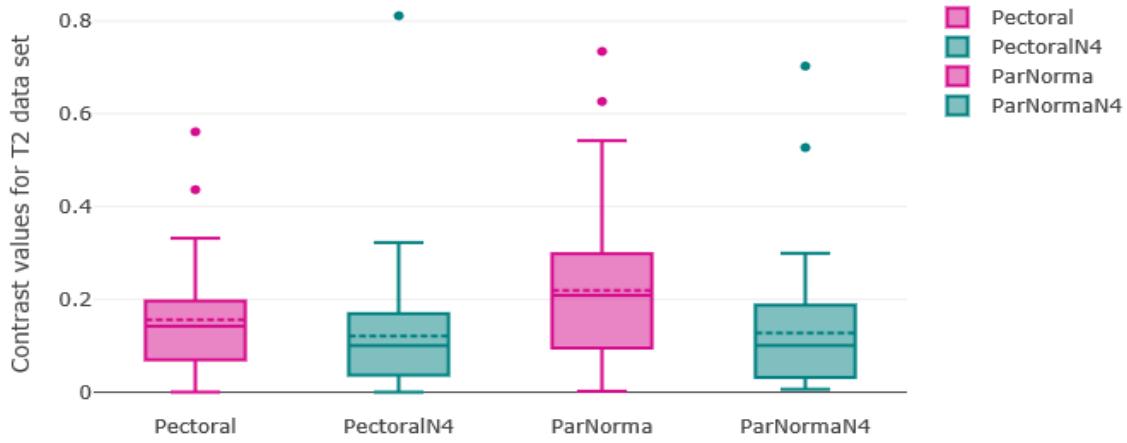


Figure 10: Left-Right contrast in the pectoral muscle and in the normal breast before and after N4ITK correction. Case of T2 weighted images

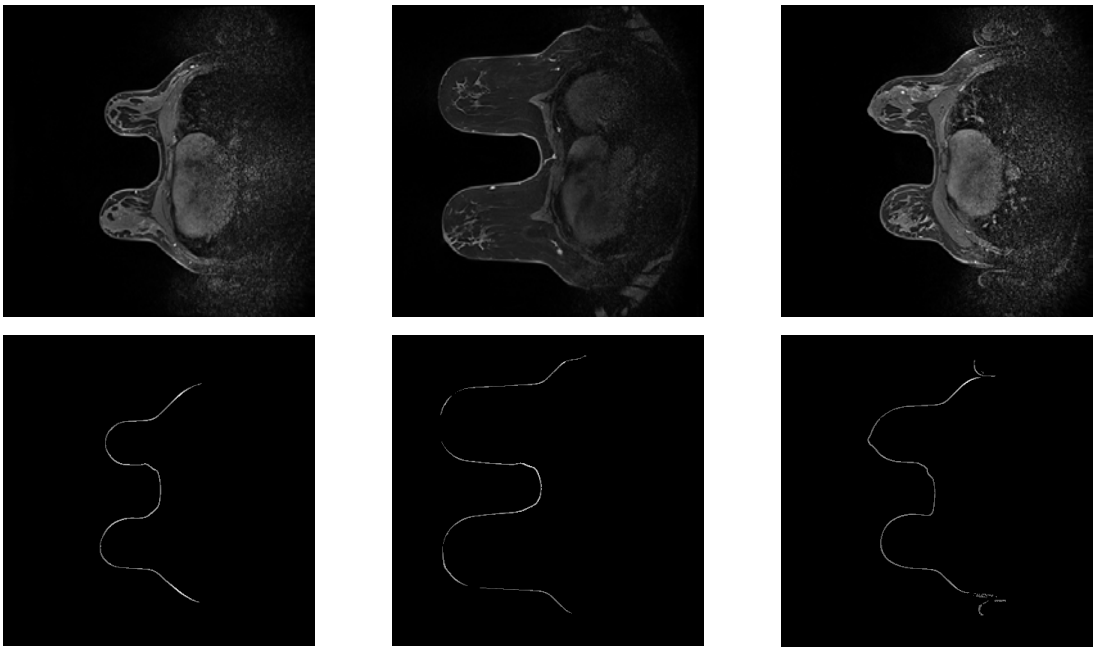


Figure 11: A few examples displaying the robustness of the pipeline

similarities in the shape of the histogram curve for some specific tissues, such as subcutaneous fat. We notice a trend in the histograms of the scans that were acquired using the same coil. Histograms of subcutaneous fat from scans acquired with the *constructor breast coil* display an initial distinct peak (see Fig. [12]), whereas scans acquired with the *Sentinelle biopsy coil* display a histogram with several peaks stretched out in a plateau form (see Fig. [13]). In most cases the highest peak is in the center of the plateau.
To visualize the voxels corresponding to the first peak of the histogram, we reconstruct the voxels corresponding to the first peak plus and minus 1%. In scans acquired with the *constructor breast coil*, we notice that most of the subcutaneous fat voxels are displayed. In scans acquired using *Sentinelle biopsy coil*, we noticed fewer points in the reconstructed image.

5. Discussion

Our study aims at tackling the challenges facing breast MR standardization. We successfully explored the hyper parameters of the N4 algorithm to achieve the best results in the shortest possible time, since homogeneity correction is an important post-processing step for MR breast analysis. Furthermore, we found a self defined mask of the breast region proved to be quite efficient in non-homogeneity correction. From the literature review related to MR brain standardization, methods based on histogram analysis have shown to be quite successful. We hoped to check histogram analysis of the subcutaneous fat layer of the breast. This layer was a good candidate for being a reference tissue in T1 DCE breast MRI, as it has high signal intensities and it corresponds to a specific region. We successfully managed to construct a pipeline for the extraction of breast subcutaneous layer. The pipeline proved to achieve its goal for the 44 MR volumes of our data set.
In light of investigating our hypothesis regarding histogram based breast tissue standardization, we discovered the similarity in histogram shape in scans acquired with the same MR coil. Scans acquired using the constructor coil show a single and distinct peak in the subcutaneous fat whereas scans using the Sentinelle biopsy coil display a more spread out plateau shaped histogram with more than one peak. The discovery of the different histogram pattern opens the way to test various other methods that are mentioned in the state of the art section. It will also be interesting to explore the reasons for the differences in results acquired with the two coils, some phantoms studies are scheduled to further investigate that point.

6. Conclusions

In conclusion, our study demonstrates the importance of non homogeneity correction for breast MR scans. Due to the double coil nature of breast MR acquisition, it may lead to a significant amount of bias in the scans. Hence if default parameters for N4 bias correction are efficient for brain MR images, they are not well-suited for breast scans. The study also presents a computationally convenient segmentation pipeline for the extraction of breast subcutaneous fat layer. Further analysis of the histogram shape of this fat layer revealed an unknown confounding factor in MR breast analysis, due to the

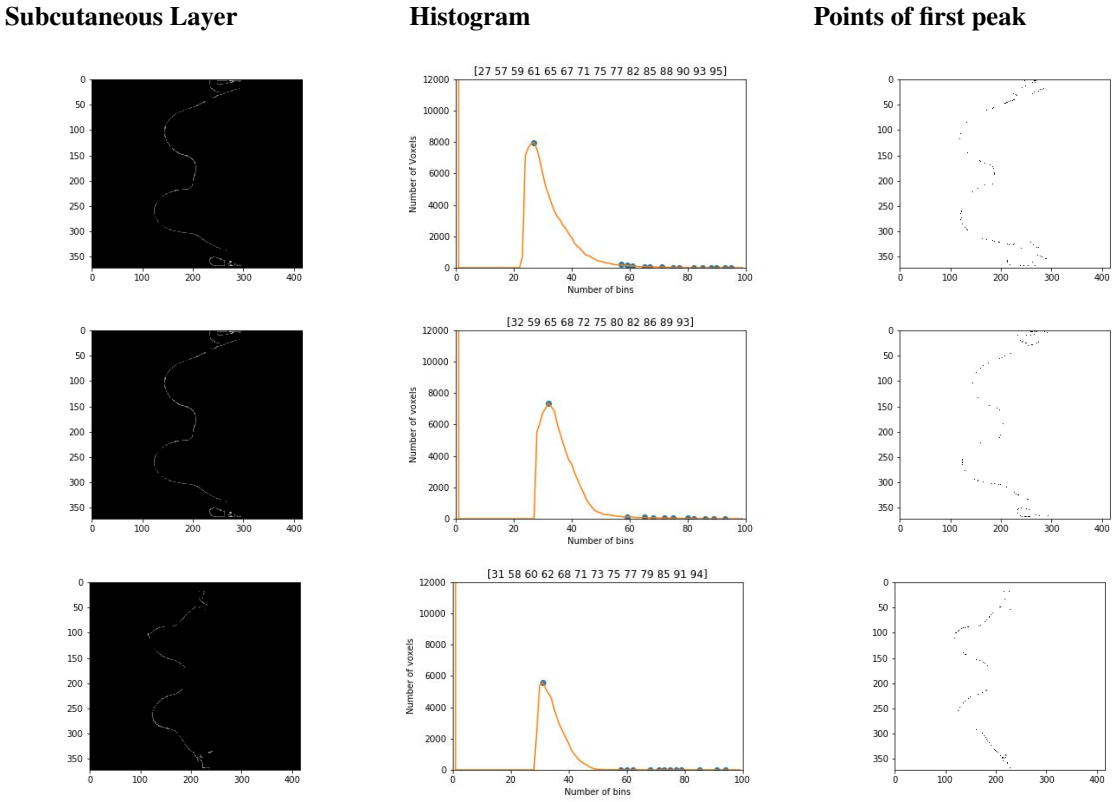


Figure 12: A few examples of patients acquired with the *constructor breast coil*. From left to right: the subcutaneous fat, the histograms of image intensities in this layer, the points corresponding to a small window centred around the first peak of the histogram.

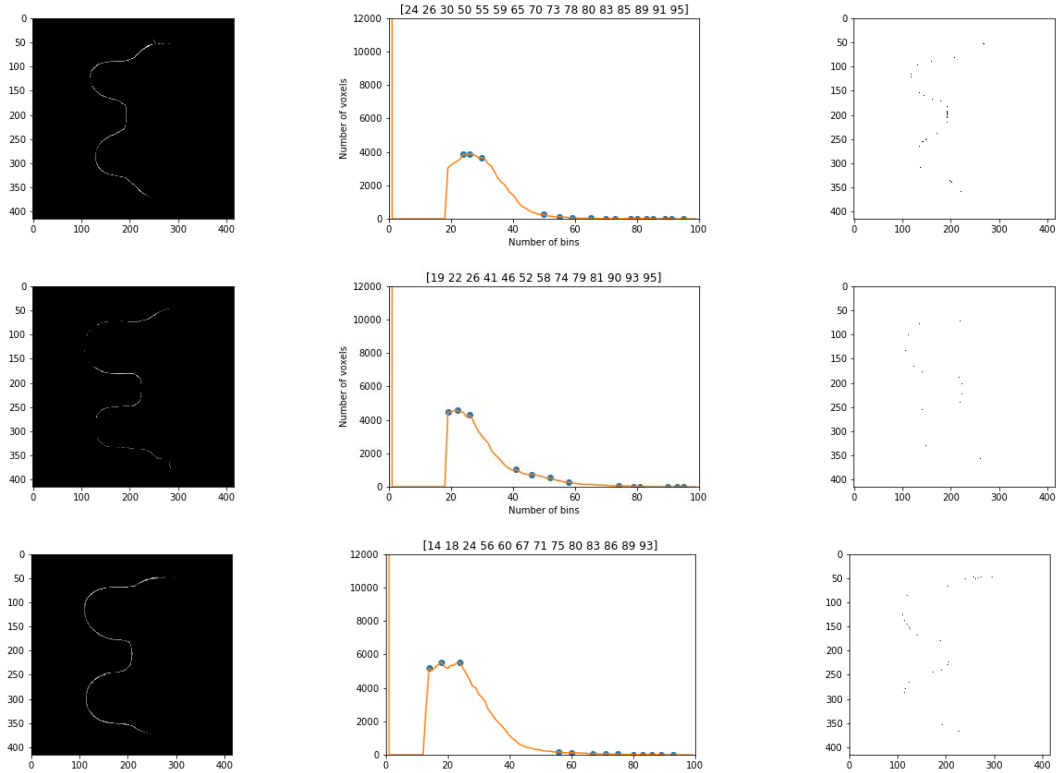


Figure 13: A few examples of patients acquired with the *Sentinelle biopsy coil*. From left to right: the subcutaneous fat, the histograms of image intensities in this layer, the points corresponding to a small window centred around the first peak of the histogram.

different coils used for MR scan. It gives rise to further investigate the effects of the two coils on the MR signal intensities inside the breast.

7. Acknowledgments

The project was done as a master thesis project at IMIV Laboratory, Inserm-CEA, Service Hospitalier Frederic Joliot, Orsay, France in collaboration with Institut Curie, Paris where the MR scans were acquired. We thank *Dr Irene Buvat*, head of the IMIV laboratory, for her availability and her valuable advice. We are indebted to *Dr Caroline Malhaire* and *Dr Herve Brisse* from the Radiology Department at Institut Curie for providing us MR data sets. We are grateful to *Dr Pia Akl*, from Institut Curie, the expert radiologist who helped us in validating this work and we thank her for the fruitful discussions we could have together.

8. References

, . American Cancer Society breast cancer facts figures. `https://www.cancer.org/research/cancer-facts-statistics/breast-cancer-facts-figures.html`. Accessed: 2019-06-14.

Aerts, H.J., 2016. The potential of radiomic-based phenotyping in precision medicine: a review. JAMA oncology 2, 1636–1642.

Ahmed, M.N., Yamany, S.M., Mohamed, N., Farag, A.A., Moriarty, T., 2002. A modified fuzzy c-means algorithm for bias field estimation and segmentation of mri data. IEEE transactions on medical imaging 21, 193–199.

Belli, P., Costantini, M., Malaspina, C., Magistrelli, A., Latorre, G., Bonomo, L., 2006. Mri accuracy in residual disease evaluation in breast cancer patients treated with neoadjuvant chemotherapy. Clinical radiology 61, 946–953.

Brasic, N., Wisner, D.J., Joe, B.N., 2013. Breast mr imaging for extent of disease assessment in patients with newly diagnosed breast cancer. Magnetic Resonance Imaging Clinics 21, 519–532.

Christensen, J.D., 2003. Normalization of brain magnetic resonance images using histogram even-order derivative analysis. Magnetic resonance imaging 21, 817–820.

Delille, J.P., Slanetz, P.J., Yeh, E.D., Halpern, E.F., Kopans, D.B., Garrido, L., 2003. Invasive ductal breast carcinoma response to neoadjuvant chemotherapy: noninvasive monitoring with functional mr imagingâĂŤpilot study. Radiology 228, 63–69.

Dialani, V., Chadashvili, T., Slanetz, P.J., 2015. Role of imaging in neoadjuvant therapy for breast cancer. Annals of surgical oncology 22, 1416–1424.

Ertas, G., Doran, S.J., Leach, M.O., 2017. A computerized volumetric segmentation method applicable to multi-centre mri data to support computer-aided breast tissue analysis, density assessment and lesion localization. Medical & biological engineering & computing 55, 57–68.

Ertaş, G., Gülçür, H.Ö., Osman, O., Uçan, O.N., Tunacı, M., Dursun, M., 2008. Breast mr segmentation and lesion detection with cellular neural networks and 3d template matching. Computers in biology and medicine 38, 116–126.

Gevaert, O., Mitchell, L.A., Achrol, A.S., Xu, J., Echegaray, S., Steinberg, G.K., Cheshier, S.H., Napel, S., Zaharchuk, G., Plevritis, S.K., 2014. Glioblastoma multiforme: exploratory radiogenomic analysis by using quantitative image features. Radiology 273, 168–174.

Gillies, R.J., Kinahan, P.E., Hricak, H., 2015. Radiomics: images are more than pictures, they are data. Radiology 278, 563–577.

Giordano, S.H., 2003. Update on locally advanced breast cancer. The oncologist 8, 521–530.

Goya-Outi, J., Orlhac, F., Calmon, R., Alentorn, A., Nioche, C., Philippe, C., Puget, S., Boddaert, N., Buvat, I., Grill, J., et al., 2018. Computation of reliable textural indices from multimodal brain mri: suggestions based on a study of patients with diffuse intrinsic pontine glioma. Physics in Medicine & Biology 63, 105003.

Hayton, P., Brady, M., Tarassenko, L., Moore, N., 1997. Analysis of dynamic mr breast images using a model of contrast enhancement. Medical image analysis 1, 207–224.

Held, K., Kops, E.R., Krause, B.J., Wells, W.M., Kikinis, R., Muller-Gartner, H.W., 1997. Markov random field segmentation of brain mr images. IEEE transactions on medical imaging 16, 878–886.

Jager, F., Hornegger, J., 2008. Nonrigid registration of joint histograms for intensity standardization in magnetic resonance imaging. IEEE Transactions on Medical Imaging 28, 137–150.

Knuttel, F.M., Menezes, G.L., van den Bosch, M.A., Gilhuijs, K.G., Peters, N.H., 2014. Current clinical indications for magnetic resonance imaging of the breast. Journal of surgical oncology 110, 26–31.

Kong, X., Moran, M.S., Zhang, N., Haffty, B., Yang, Q., 2011. Meta-analysis confirms achieving pathological complete response after neoadjuvant chemotherapy predicts favourable prognosis for breast cancer patients. European journal of cancer 47, 2084–2090.

Kumar, V., Gu, Y., Basu, S., Berglund, A., Eschrich, S.A., Schabath, M.B., Forster, K., Aerts, H.J., Dekker, A., Fenstermacher, D., et al., 2012. Radiomics: the process and the challenges. Magnetic resonance imaging 30, 1234–1248.

Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., Van Stiphout, R.G., Granton, P., Zegers, C.M., Gillies, R., Boellard, R., Dekker, A., et al., 2012. Radiomics: extracting more information from medical images using advanced feature analysis. European journal of cancer 48, 441–446.

Larsen, C.T., Iglesias, J.E., Van Leemput, K., 2014. N3 bias field correction explained as a bayesian modeling method, in: Bayesian and graphical models for biomedical imaging. Springer, pp. 1–12.

Lee, S., Wolberg, G., Shin, S.Y., 1997. Scattered data interpolation with multilevel b-splines. IEEE transactions on visualization and computer graphics 3, 228–244.

Leung, K.K., Clarkson, M.J., Bartlett, J.W., Clegg, S., Jack Jr, C.R., Weiner, M.W., Fox, N.C., Ourselin, S., Initiative, A.D.N., et al., 2010. Robust atrophy rate measurement in alzheimer's disease using multi-site serial mri: tissue-specific intensity normalization and parameter selection. Neuroimage 50, 516–523.

Luangdilok, S., Samarnthai, N., Korphaisarn, K., 2014. Association between pathological complete response and outcome following neoadjuvant chemotherapy in locally advanced breast cancer patients. Journal of breast cancer 17, 376–385.

Madabhushi, A., Udupa, J.K., Moonis, G., 2006. Comparing mr image intensity standardization against tissue characterizability of magnetization transfer ratio imaging. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine 24, 667–675.

Marinovich, M., Sardanelli, F., Ciatto, S., Mamounas, E., Brennan, M., Macaskill, P., Irwig, L., Von Minckwitz, G., Houssami, N., 2012. Early prediction of pathologic response to neoadjuvant therapy in breast cancer: systematic review of the accuracy of mri. The Breast 21, 669–677.

Padhani, A.R., Hayes, C., Assersohn, L., Powles, T., Makris, A., Suckling, J., Leach, M.O., Husband, J.E., 2006. Prediction of clinicopathologic response of breast cancer to primary chemotherapy at contrast-enhanced mr imaging: initial clinical results. Radiology 239, 361–374.

Parekh, V., Jacobs, M.A., 2016. Radiomics: a new application from established techniques. Expert review of precision medicine and drug development 1, 207–226.

Robitaille, N., Mouiha, A., Crépeault, B., Valdivia, F., Duchesne, S., 2012. Tissue-based mri intensity standardization: application to multicentric datasets. Journal of Biomedical Imaging 2012, 4.

Shah, M., Xiao, Y., Subbanna, N., Francis, S., Arnold, D.L., Collins, D.L., Arbel, T., 2011. Evaluating intensity normalization on mris of human brain with multiple sclerosis. Medical image analysis 15, 267–282.

Shin, H., Kim, H., Ahn, J., Kim, S., Jung, K., Gong, G., Son, B., Ahn, S., 2011. Comparison of mammography, sonography, mri and clinical examination in patients with locally advanced or inflammatory breast cancer who underwent neoadjuvant chemotherapy. The British journal of radiology 84, 612–620.

Thompson, A., Moulder-Thompson, S., 2012. Neoadjuvant treatment of breast cancer. annals of Oncology 23, x231–x236.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C., 2010. N4itk: improved n3 bias correction. IEEE transactions on medical imaging 29, 1310.

Tustison, N.J., Gee, J.C., 2005. Nd ck b-spline scattered data approximation. The Insight Journal 2, 88.

Tustison, N.J., Gee, J.C., 2006. Generalized n-d c k b-spline scattered data approximation with confidence values, in: International Workshop on Medical Imaging and Virtual Reality, Springer. pp. 76–83.

Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 1999. Automated model-based tissue classification of mr images of the brain. IEEE transactions on medical imaging 18, 897–908.

Weisenfeld, N.I., Warfield, S.K., 2004. Normalization of joint image-intensity statistics in mri using the kullback-leibler divergence., in: ISBI, pp. 101–104.

Wells, W.M., Grimson, W.E.L., Kikinis, R., Jolesz, F.A., 1996. Adaptive segmentation of mri data. IEEE transactions on medical imaging 15, 429–442.

Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. IEEE transactions on medical imaging 20, 45–57.

Medical Imaging and Applications

Master Thesis, June 2019

Improving Generalization of Convolution Neural Networks for Digital Pathology by Minimizing Stain Heterogeneity through Normalization, Augmentation and Domain Learning

Amjad Khan, Henning Müller

Institute of Information Systems
University of Applied Sciences Western Switzerland (HES-SO)
3960-Sierre, Switzerland

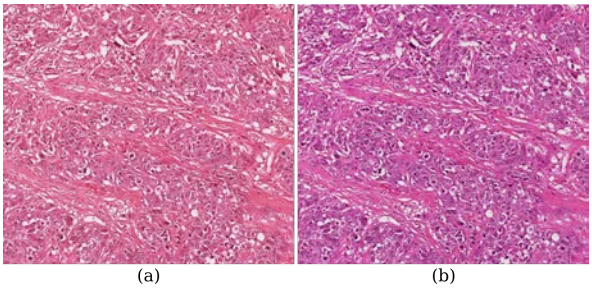
Abstract

Histopathology is the gold standard to diagnose and grade various kinds of cancer. However, stain color heterogeneity can exist in histopathology slides of the same tissue type prepared in different pathology institutes due to various factors arising from the staining procedure and environment. The stain color heterogeneity further effects the generalization performance in machine learning based computational analysis of digital pathology. It is challenging for a computer aided diagnostic system to produce identical results on the slides prepared at different histology settings with the same tissue and diagnostic objectives. In this thesis, the stain color heterogeneity across various pathology centers has been investigated to minimize its effects on a convolution neural network based classification problem through various preprocessing and domain adaptation methods. Several public histopathology databases have been explored to reach at a suitable collection to conduct experiments on stain color heterogeneity. By considering the suitable dataset, various stain color normalization and augmentation techniques are quantified on tumor and normal tissue classification task to improve the generalization on external data. In addition to the stain color normalization and augmentation techniques, the convolution neural network is also trained to learn the domain information of the samples while training for class label classification. Comparative analysis of obtained results have shown significance performance of the classifier on an external data when trained with stain normalization and augmentation methods. Top ranked methods have shown even improved results on external test samples when probability based fusion is employed.

Keywords: Digital pathology, Stain heterogeneity, Normalization, Augmentation, Domain learning, Machine learning

1. Introduction

Cancer is recognized by the World Health Organization as an important disease with global deaths amounting to 9.6 million persons in 2018 alone. 18.1 million new cases were diagnosed in the same year (WHO, 2018). Cancer is a collection of more then 100 diseases related to the uncontrolled division of body cells into the surrounding tissue (NIH, 2007). Such proliferation of abnormal cells can be malignant, integrating nearby cells into metastases. Lung, breast, colorectum and prostate are the four most frequently effected organs by cancer, so with the highest mortality rates (WHO, 2018). Histopathology is the gold standard to diagnose and grade various kinds of cancer (Khosravi et al., 2018;



(a) (b)

Figure 1: Stain color heterogeneity in digital histopathology data of the same tissue scanned with (a)Aperio and (b)Hamamatsu scanners.

Xu et al., 2017). The tissue samples of cancers are extracted through biopsies for microscopic analysis. Then these tissue samples are processed through a chemical procedure called histological staining. There are various stages involved in histological staining such as fixation, processing, embedding and sectioning (Alturkistani et al., 2015). Histological staining is an essential chemical procedure that helps to discriminate different structures within the tissue under a microscopic study. For instance, hematoxylin and eosin (H&E) are the two most frequently used chemical dyes that enhance the contrast of the tissue by coloring the cell nuclei purple and connective tissue as pink respectively. Tissue samples without staining have less contrast and appear rather grey (Roy et al., 2018). Finally, the stained slides of tissue samples can be scanned for pathologists and researchers to analyze them further in a digital way (Pantanowitz et al., 2011). Stain color heterogeneity exists in whole slide images (WSIs) from distinct laboratories (see Figure 1) due to the factors involved in staining and the scanning process (Yagi, 2011). Variations in thickness of specimen, staining and properties of digital scanners make it inviable to obtain homogeneous data from multiple laboratories (Tellez et al., 2019; Yagi, 2011). Heterogeneity in WSIs of the same tissue type acquired from different laboratories or scanners makes it challenging for machine learning, mainly deep learning based approaches to generalize identical results in computer aided diagnosis across all WSIs (Zheng et al., 2019).

This master thesis investigates the staining heterogeneity across public histopathology data-set of the same tissue, taken from distinct staining laboratories or acquired from different scanners. The main objectives of the research work are as following,

- Investigation of various public histopathology data-sets of different tissue type that include Colon, Lung, Prostate and Breast to reach a data that can be further utilized in stain color heterogeneity study.

- Investigating various stain color normalization and augmentation approaches that address stain heterogeneity on the suitable collection with diagnostics tasks such as cancer classification.

- Combining the best performing stain color normalization and augmentation methods to improve the generalization on the external data-set by using a convolution neural network based approach.

- Evaluate the performance of the convolution neural network by learning domain heterogeneity of the corresponding training samples while class label classification.

2. State of the art

2.1. Reference Data-sets for Stain Heterogeneity

In order to conduct experimental analysis to quantify generalization performance under heterogeneous conditions, various public histopathology data-sets were investigated, as shown in Table 1. These collections are from various tissue types, such as colon, lung and bronchus, prostate, breast and ovary. The selection of the data-sets was limited to H&E stained whole slide images (WSIs), patches and tissue micro arrays (TMA). During data exploration and search, 18 public data sets were found from several resources. Theses resources include public research repositories, grand challenges and individual repositories of the researchers or institutes. In the following paragraphs the characteristics of each data set is briefly described. The Cancer Genome Atlas (TCGA) is largest online repository for cancer genomics and molecular characterization. It is initiated by National Cancer Institute and the National Human Genome Research Institute in 2006, now it is extended to multiple institutions. The data is publicly available for diagnosis, treatment and prevention of cancer research. TCGA contained three histopathology datasets of our interest, TCGA-COAD, TCGA LSCC & LUAD and TCGA-PRAD, all of them are the collections from multiple centers and annotated globally for different types and subtyes of cancer. TCGA-COAD is a collection of 4 types of colon cancer and consisted of 453 WSIs, digitized at 40× magnification with an average resolution of $100,000 \times 100,000$ pixels. TCGA-COAD is public storage with controlled access that requires a specific registration process to access the data according the data access policy. TCGA LSCC & LUAD contains 956 WSIs of lung and bronchus with the same resolution, magnification and access type as TCGA-COAD. Whereas TCGA-PRAD is an open access prostate adenocarcinoma data and available in 272 WSIs from 25 different institutes with global annotations of Gleason grades (Arvaniti et al., 2018). Tissue microarrays from University Hospital Zurich (TMA-Zurich) is also public repository that contains 71 prostate cases scanned by Hamamatsu(C9600 NanoZoomer 2.0-HT) slide scanner at a magnification and a resolution of 40× and $7,000 \times 7,000$ pixels respectively (Arvaniti et al., 2018; Zhong et al., 2017).

Automatic cancer detection and classification in whole slide lung histopathology (ACDC@LUNGHP) is a challenge providing 200 WSIs, digitized at First Hospital of Changsha, China by Olympus VS120 with 20× magnification (Li et al., 2018). ACDC@LUNGHP contains the local annotations of cancer regions and publicly available to the registered participant in the challenge (Li et al., 2018). Kather et al. (2016) collected 5000 patches of 150×150 pixels with a magnification of 20× at Institute of Pathology, Heidelberg University.

Table 1: List of the investigated histopathology data-sets of various tissue types for stain color heterogeneity study across the histology centers.

Dataset	Tissue	Images (Type)	Center	Scanner
TCGA-COAD[1]	Colon	453 (WSIs)	Multiple centers	-
CRCHistoPhenotypes[2]	Colon	200 (Patches)	University Hospitals Coventry and Warwickshire, UK	Omnyx VL120
GIaS16_warwick_QU[3]	Colon	165(Patches)	University Hospitals Coventry and Warwickshire, UK	Zeiss MIRAX MIDI
CRAG_v1[4]	Colon	213 (Patches)	University Hospitals Coventry and Warwickshire, UK	Omnyx VL120
Kether Textures[5]	Colon	5000 (Patches)	Institute of Pathology,Heidelberg University, Mannheim, Germany	Aperio ScanScope
ACDC@LUNGHP[6]	Lung	200 (WSIs)	First Hospital of Changsha, China	Olympus VS120
TCGA LSCC and LUAD[1]	Lung	956 (WSIs)	Multiple centers	-
Stanford TMA database[7]	Lung	Multiple (TMAs)	School of Medicine, Stanford University, USA	-
TCGA-PRAD[1]	Prostate	272 (WSIs)	25 centers	-
TMA Zurich[8]	Porstate	71 (TMAs)	University Hospital Zurich	Hamamatsu (C9600 NanoZoomer 2.0-HT)
TUPAC[9]	Breast	821 (WSIs)	Multiple centers	-
PatchCamelyon[10]	Breast	327680 (Patches)	Radboud University Medical Center, University Medical Center Utrecht, Netherlands	3DHistech Pannoramic Flash II 250, Hamamatsu NanoZoomer-XR C12000-01
UCSB dataset[11]	Breast	58 (Patches)	Center for Bio-image Informatics, USA	-
MITOS-ATYPIA-14[12]	Breast	1420 (Patches)	Pathology Department at Piti-Salpłtrire Hospital in Paris, France	Aperio Scanscope,XT Hamamatsu, Nanozoomer 2.0-HT
Camelyon 16[13]	Breast	399 (WSIs)	Radboud University Medical Center, University Medical Center Utrecht, Netherlands	3DHistech Pannoramic Flash II 250, Hamamatsu NanoZoomer-XR C12000-01
Camelyon 17[14]	Breast	1000 (WSIs)	Radboud University Medical Center, Canisius-Wilhelmina Hospital, University Medical Center Utrecht, Rijnstate Hospital, and Laboratorium Pathologie Oost-Nederland, Netherlands	3DHistech Pannoramic Flash II 250, Hamamatsu NanoZoomer-XR C12000-01, Philips Ultrafast Scanner
Lymphoma[15]	Lymphoma	256 (Patches)	Intramural Research Program Laboratory of Genetics, National Cancer Institute, National Institute on Aging	-
SFU dataset[16]	Ovary	133 (WSIs)	6 different pathology centers, Canada	Aperio ScanScope

Note: WSIs: Whole Slide Images, TMAs: Tissue Microarrays

[1] https://portal.gdc.cancer.gov/, [2] https://warwick.ac.uk/fac/sci/dcs/research/tia/data/crchistolabelednucleihe/, [3] https://warwick.ac.uk/fac/sci/dcs/research/tia/glascontest, [4] https://warwick.ac.uk/fac/sci/dcs/research/tia/data/mildnet/
[5] https://zenodo.org/record/53169#.XHzqdN1KjRY, [6] https://acdc-lunghp.grand-challenge.org/, [7] https://tma.im/cgi-bin/viewArrayBlockList.pl, [8] http://dx.doi.org/10.7910/DVN/4WEMEQ, [9] http://tupac.tue-image.nl/
[10] https://github.com/basveeling/pcam, [11] https://bioimage.ucsb.edu/research/bio-segmentation, [12] https://mitos-atypia-14.grand-challenge.org/Dataset/, [13] https://drive.google.com/drive/folders/0BzsdkU4jWx9Bb19WNndQTlUwb2M
[14] https://drive.google.com/drive/folders/0BzsdkU4jWx9BaXVHSXRJTnpLZU0, [15] https://ome.grc.nia.nih.gov/iicbu2008/, [16] http://ensc-mica-www02.ensc.sfu.ca/download/

The collection is consisting of 8 different textures in colorectal low grade and high grade tumors. The patches were annotation manually and collection is available publicly. Similarly, for colon cancer research three public datasets acquired at Department of Pathology, University Hospitals Coventry and Warwickshire. First two collections, CRCHistoPhenotypes and CRAG v1 contain 200 and 213 patches of 500×500 and 1500×1500 pixels respectively (Awan et al., 2017). These datasets are digitized with OmnyxVL120 and locally annotated for nuclie segmentation and colorectal adenocarcinoma grading. GIaS16_warwick_QU is the third collection that comes with 165 patches of 775×522 pixels with $20\times$ magnification, digitzied with Zeiss MIRAX MIDI scanner and annotated for gland segmentation task (Sirinukunwattana et al., 2016). Stanford tissue microarray database is an other large collection of histopathology images of various organs aggregated by School of Medicine, Stanford University (Marinelli et al., 2008). Multiple TMAs are availabe of human lung with global annotations of malignant and benignant and data can be available on request. BenTaieb et al. (2017) aggregated

very well structured collection of ovarian carcinoma subtyping of 133 patients, slides were digitized by Aperio ScanScope with a highest magnification of 40x and having an average resolution of $50,000 \times 50,000$ pixels. WSIs were annotated by a common agreement of three expert pathologists for carcinoma subtyping and available publicly. The National Institute of Aging provides digitized histopathology images of various organs of human and animal as benchmark for testing and comparing the performance of the different image analysis algorithms (Shamir et al., 2008).

Moreover, numerous collections of histopathology images from women breast cancer were studied. Tumor proliferation assessment challenge (TUPAC) is one of the large collections of woman breast cancer research, consisting of 821 WSIs acquired from multiple centers and annotated for tumor regions (Veta et al., 2018). Similarly, Center for Bio-image Informatics is an online repository containing various bioblogical datasets for benchmarking of segmentation and classification (Drelie Gelasca et al., 2008). Among other datasets, it provides UCSB dataset which consists of 58 patches of two different resolutions 896×768 and 768×512 pixels with a global annotations of breast cancer malignant and benignant cases. MITOS-ATYPIA is also an other interesting collection of breast cancer images by Pathology Department at Piti-Salpłtrire Hospital in Paris (Roux et al., 2014). In the collection 1420 patches are from two different resolutions and magnifications $1539 \times 1376, 1663 \times 1485$ pixels and $20\times, 40\times$ respectively. The same tissues were digitized with two different slide scanners, Aperio Scanscope XT, Hamamatsu Nanozoomer 2.0-HT and color heterogeneity can be visualized clearly. The patches were annotated for scoring nuclear atypia and publicly available. In 2016, Camelyon16 challenge took place about the diagnostic assessment of deep learning methods for lymph node metastases detection in women suffering with breast cancer (Veta et al., 2018). It provides 399 WSIs from two hospitals in Netherlands, Radboud University Medical Center and University Medical Center Utrecht. The WSIs were acquired from two different scanners 3DHISTECH: Pannoramic 250 Flash II and Hamamatsu: NanoZoomer-XR with $20\times$ and $40\times$ magnifications. The slides were annotated by the pathologists for macrometastases and micrometastases. Although the challenge is over, however, dataset is still available to download from the public repository (Litjens et al., 2018). PatchCamelyon is an other public repository which is an extraction of patches from Camelyon16 WSIs to benchmark deep learning algorithms for tumor classification task (Veeling et al., 2018). It provides well organized train, validation and test sets all consisting of $327,680$ patches with a resolution of 96×96 pixels. The patches were extracted so that each positive label indicates that the center region of 32×32 pixels in a patch contains at least one pixel from tumor tissue. Finally, Camelyon17 challenge is also having a repository of breast cancer WSIs which is the extension of Camelyon16. It provides 1000 WSIs from 5 different centers (Litjens et al., 2018). These centers are from Netherlands and include: Radboud University Medical Center, Canisius-Wilhelmina Hospital, University Medical Center Utrecht, Rijnstate Hospital, and Laboratorium Pathologie Oost-Nederland. The slides were locally annotated for metastases and well organized training set contains WSIs from each center. From all above collections, It is concluded that the characteristics of Camelyon17 training set are suitable for the color heterogeneity experimentation. Therefore, in this thesis, 50 lesion-level annotated slides that provide 10 slides from each center are used to quantify different methods for color heterogeneity that exits due to the slide preparation method and scanning of histopathology slides at five different pathology centers.

2.2. Global Stain Color Normalization

In order to overcome stain color heterogeneity, various studies have been conducted to homogenize the data acquired from different laboratories or different scanners of the same tissue type with a common goal of diagnosis. It is worth mentioning that the stain heterogeneity in digital histopathology does not only effects the computational analysis performed by the research community but it is considered as a problem for the pathologists when they perform visual analysis on the whole slide images especially in telepathology. Such a problem was highlighted by Yagi (2011), where extensive experiments were performed to calibrate display systems using different color filters for H&E staining. In this study, staining, thickness of specimen, scanner and scanning process, viewing software and displaying systems were considered responsible for color heterogeneity. The display systems study through Macbeth color chart was helpful to calibrate the color of the displays across entire department to make the visual effects homogeneous in histopathology data.

Stain normalization in histopathology is not new to the digital image processing domain especially when it comes to color variation in the image due to incandescent illumination, where it is important to bring different images of the same scene to a common color distribution. Identification of such problem was presented by Reinhard et al. (2001) and it was suggested to transfer the color distribution of data to the color characteristics of one common reference image among the data. The experimental results presented by Reinhard et al. (2001) were mostly focused on outdoor scenes, however, due to simplicity of the method, it has been utilized for histopathology slides normalization. The Reinhard et al. (2001) method is based on color distribution model in each channel of the *Lab* color space. In histopathlogy, the digitizing process of histology slides is quite similar to image acquisition in other various

applications, however, digitized slides contain variable microscopic information depending on the magnification used. In the slide scanning process, the appearance of the stain depends on the intensity absorbed by the tissue and it further depends on amounts of the stain added to the tissue and its handling and storage methods. The linear relationship between the stains and absorbed intensity in the tissue is described by the corresponding optical density for stain normalization using devolution model(Macenko et al., 2009). In the method, the stain colors were estimated using a singular value decomposition (SVD) based matrices and through linear per channel normalization based on the 99th percentile to map source image values to the target matching. However, the Macenko et al. (2009) method depicts inconsistencies in the performance if a slide under normalization contains higher number of stains. Furthermore, the method modifies the color distributon of both source and target images which is undesirable in some cases where suitable reference image is used to map the characteristics to the rest of data. In order to overcome the problems faced by method presented by Macenko et al. (2009); Reinhard et al. (2001), a non linear mapping of channel statistics based normalization method from source to target images is introduced by Khan et al. (2014). The method was mainly focused on estimated stain matrix, color deconvolution and reconstruction steps. The stain matrix estimation was performed by color classification. In the color classification task a Relevance Vector Machine (RVM) was trained in RGB color model. Significance of the method depends on the robust deconvoulution matrix estimation and mapping function. The color deconvolution separates out the variation of each stain to correct it independently. However, the pre-trained RVM color classifier makes the method unstable in the test cases that deviate from the train cases due to varying dye color. Moreover, the manual or random choice of the target image in the absence of the prior stain and biological information in it make the normalization task to expect random generalization on certain applications.

Ehteshami Bejnordi et al. (2016) proposed a color standardizing technique for whole slide image by using the color and spatial information to classify the pixels into stain components. The density and chromatic distribution of the data in the hue-saturation-density color space is aligned with a template slide. However, the performance of the method in the new data relies on expert opinion about the chosen reference template slide by considering the color and cellular information in to account. On the other hand, Tam et al. (2016) illustrated contrast limited adaptive histogram equalization(CLAHE) based intensity centering model to bring the color distribution to the center point within whole data. The method avoids the reference and target image significant statistics, however, the histogram equalization is limited to the spatial dependency of pixels. In order to preserve the biological structural information while performing the color noramalizatin taks,Vahadane et al. (2016) presented the structure preserving color normalization. In the method, the stain density map was considered as sparse and non-negative. In the sparsity it was assumed that the biological material occupies one given pixels location or other but not both. Similarly, the non-negativity describes that either a biological material is absorbing the light or not and optical density cannot be negative. Based on the above assumptions the color appearance and stained density matrices for both source and target images were estimated for color transformation.

2.3. Machine Learning based Stain Normalization

The stain color normalization techniques are able to cope with the variability of the stain and appearance of the digital histopathology for visual observations. The stain heterogeneity is also dealt with several machine learning based approaches. These techniques focus on generalization improvement in computational analysis by considering the texture feature along with color information during normalization process. In this context, deep convolution feature-aware normalization was presented by Bug et al. (2017). The study mainly focused on the visually relevant image deep features and style transferring. The feature aware normalization was inspired by batch normalization (Ioffe and Szegedy, 2015) and long term memory (Hochreiter and Schmidhuber, 1997) mechanisms. The method performed pixel wise transformation based on features contained in the plasma or nucleus in the tissue. The color was treated as a form of a style to integrate into the network by shifting and scaling parameters of batch normalization layers. The features extraction process was mainly performed by a pretrained VGG19 architecture in both reference and source images. The mean and variance metrics of reference image features were used for color normalization by shifting and scaling network parameters. Similarly Samsi et al. (2018) normalized the histopathology images by adopting the deep learning model that has been used on natural scenes colorization (Baldassarre et al.). In this study ResNet-v3 was used to extract the features from the images and then these features were fused with encoder-decoder model. The model was trained to estimate *ab* color values of *Lab* color space by minimizing the mean squared distance error between actual and estimated values.

In (Janowczyk et al., 2017), an unsupervised stain normalization method was introduced where the sparse auto-encoders were used to normalize moving images to a template image. The pixels were clustered by using k-means according to the respective tissue partitions in the saprse auto encoded feature space. Then the color distribution of each partition in the moving data was aligned with its respective color distribution of the template. Finally, the histogram equalization was performed across

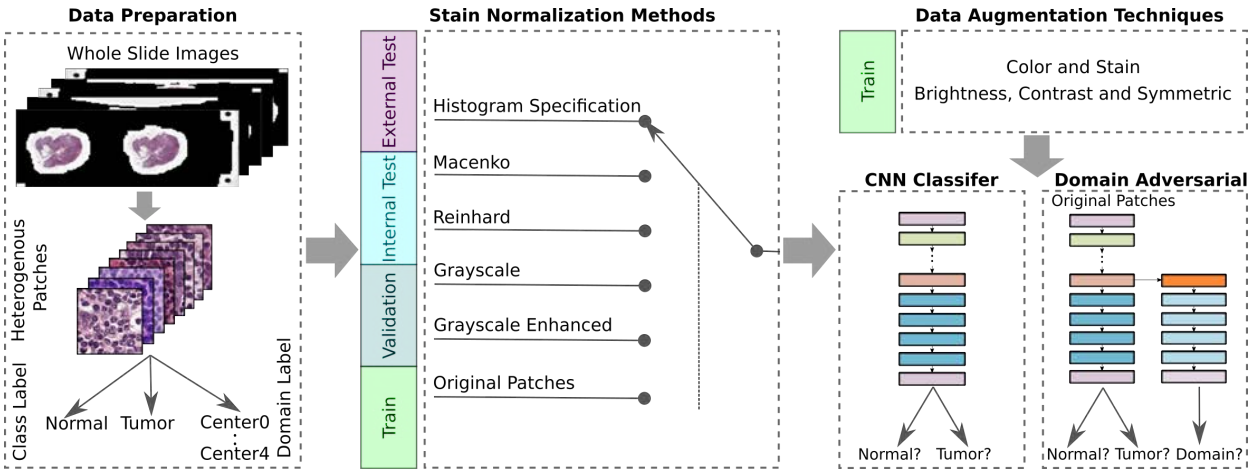


Figure 2: Block diagram representing work-flow of the system.

the color channels. The method was experimented on different data sets of various tissue types with heterogeneity due to their domains or scanners. In digital histopathology, the color normalization techniques often require the reference image to transfer the color characteristics to the rest of the images. However, Shaban et al. (2018) prosposed end-to-end generative adversarial networks (GANs) based approach to transfer the stain style by eliminating the requirement of an expert to choose a reference image. The experimentation was performed on MITOS-ATYPIA dataset which is acquired from the same tissue section with two different scanners (Roux et al., 2014). The method mainly consisted of two pairs of generator and discriminator to map the stain style of the images belonging to one domain to the other. Similarly, GANs architure was also employed by Cho et al. (2017) to transfer the stain style from source to target images. The conditional GANs were trained to learn both color distribution and histopathological patterns present in the Camelyon16 data set (Litjens et al., 2018). The images from two different centers were normalized to the gray-scale then style generator was used to colorize the gray-scale images again to a standard stain style. Inspired by style transfer and generative learning methods, (Bentaieb and Hamarneh, 2018) presented stain normalization technique by preserving the structural information. The method avoided to relay on single reference image therefore the matching was performed on stain statistics over the entire domain of images. Instead of pixel level matching, the feature representation of the images was used to normalize them. The proposed network was divided into stain transfer and task specific parts to perform both stain normalization and classification or segmentation tasks simultaneously. The stain transfer network learned the probability distribution of the images of one domain by minimizing the adversarial loss function to map the input image to a stain normalized image. Whereas the task specific model was used to maximizing the likelihood of the input image according to the given task. The proposed technique was evaluated on three different data sets of mitosis, colon and ovary (Bentaieb and Hamarneh, 2018; Roux et al., 2014; Sirinukunwattana et al., 2016) that contained color variability. Apart from stain transfer and feature based stain normalization, the generalization in convolution neural networks for computational pathology can be improved by data augmentation methods. In (Tellez et al., 2018), a data augmentation based technique was developed to improve the generalization of convolution network for histopathology data. Each patch of the train set was modified in terms of hematoxylin, eosin and residual channels then a combination of rotation, color stain, scaling, elastic deformation, image enhancement, blurring, additive Gaussian noise was used as data augmentation techniques. Extensive experimentation was conducted to improve the generalization performance on different data sets of various tissues from multiple centers (Bandi et al., 2019; Kather et al., 2016; Veta et al., 2018). Finally, a recent review (Roy et al., 2018) is considered as a reference for further analyses on various global, supervised and unsupervised color normalization methods.

3. Material and methods

In this thesis, several stain color normalization methods, data augmentation techniques and domain adversarial based learning are evaluated on a convolution neural network classifier to improve the generalization specially on the external data. The overall work-flow is shown in the Fig. 2 whereas the following subsections discuss the data, methods and tools to conduct the experiments in detail.

3.1. Data Preparation

We investigated several histopathology databases to find best suitable data that can be used for stain color heterogeneity quantification (see section 2.1). Through

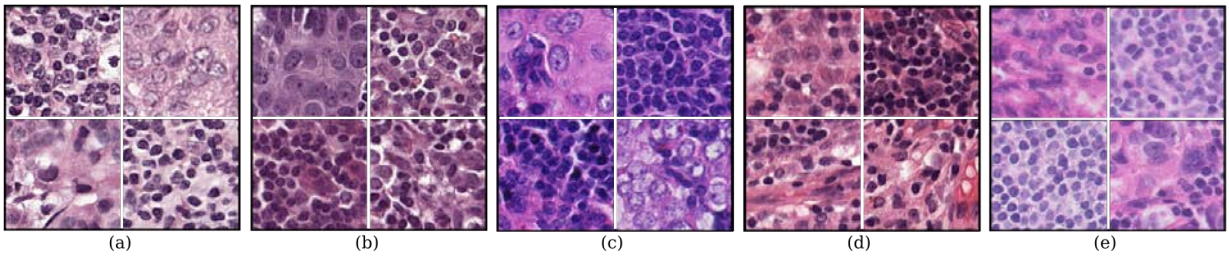


Figure 3: Tumor and normal tissue patches from whole slide images of Camelyon17 dataset (Litjens et al., 2018), where the slides were prepared and digitized at five different centers and three different scanners respectively, the color variability is clearly visible in the patches from all five centers (a to e).

this investigation, it is concluded that Camelyon17 dataset has the desirable collection to perform the experimentation. The Camelyon17 is the second grand challenge organized by Diagnostic Image Analysis Group and Department of Pathology of the Radboud University Medical Center in Netherlands. The chanllenge aimed at the evaluation of new and existing methods for detection and classification of breast cancer metastases particularly in whole slide image of histological lymph node sections. The histological slides were prepared and scanned at five different pathology centers and acquired with three different scanners (Litjens et al., 2018), where the stain color heterogeneity can be witnessed clearly as shown in Fig.3. The local annotation of tumor and normal tissues provide better way to prepare a data for convolution neural network based classification task. The slides were scanned at pixel resolution of $0.23\mu m$ to $0.25\mu m$ and provided in TIFF format. In this thesis, 50 annotated WSIs, 10 slides from each center are used to extract the patches from tumor and normal tissue areas of each slide. The tissue masks of tumor and normal regions from a 64-times down-sampled gray-scale WSI are generated by applying Otsu's method (Smith et al., 1979). Then these tissue masks are used to obtain tumor and normal regions from original RGB WSI. From each tissue region of each slide, around 500 patches of 224×224 pixels are extracted. From the tumor tissue only those patches that are covering at least 70% of tumor pixels are considered and rest are discarded. A complete process of patch extraction is highlighted in Fig.4. Finally, the extracted patches are carefully distributed to train, validation, internal test and external test so that the distribution in each partition should contain all patches obtained form a slide. In order to evaluate the performance of different augmentation and normalization techniques, five sub data sets or folds are prepared by leaving each time all patches from one center as an external test set and rest of patches from remaining four centers with slightly imbalance classes are distributed among train (70%), validation (15%) and internal test (15%) sets as shown in Table 2. It is worth mentioning that the outcomes of our experimentation are not comparable with the results of the existing techniques in the grand challenge due the different nature of experimentation. In our experimental work only those WSIs are used where the relevant center information is provided. In such case, the slides in original train set of the challenge is center-wise organized whereas the original test set does not contain such information. Therefore, the patch extraction is performed on the WSIs from the original train set of the challenge. The aim of the experimentation is to improve the generalization on the data set which should be from an external source or center then the one used in training, validation and internal testing phase with the known source.

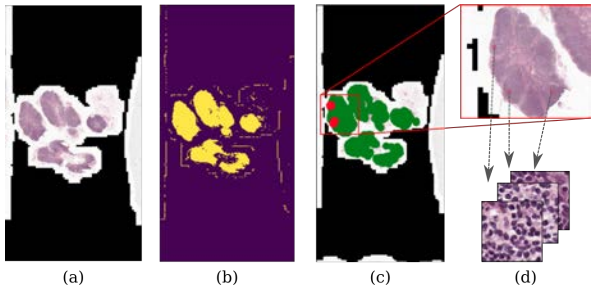


Figure 4: Patch extraction process, (a) Whole slide image, (b) segmented tissue mask, (c) annotated tumor lesions (red) and normal tissue (green) and (d) extracted patches.

Table 2: Data distribution into five different subsets or folds where each subset has external test set of patches belonging to a center and train, validation and internal sets are consisting of patches from rest of four centers.

D.Fold	Train-Validation-Internal Test Sets				E.Test Set
1	Center1	Center2	Center3	Center4	Center0
2	Center0	Center2	Center3	Center4	Center1
3	Center0	Center1	Center3	Center4	Center2
4	Center0	Center1	Center2	Center4	Center3
5	Center0	Center1	Center2	Center3	Center4

Note: D: Data, E: External

3.2. Data Augmentation

The convolution neural networks (CNNs) consist of large number of trainable parameters. Large quantity

of samples in the training process make the networks more robust by learning more features which further improves the generalization on the test samples. Data augmentation is very useful in the cases where less number of training samples are available or in imbalance class scenario where CNNs are more prone to converge towards to a class with higher number of training samples. However, in our case the data augmentation enhances the characteristics of data to improve the generalization on the external test samples that are belonging to a different source with different variations in the staining. Therefore, it is hypothesized that training a network on the train data with more staining or color variations through augmentation could improve the generalization performance by learning variability. Various data augmentation techniques are individually evaluated on training samples and then effective techniques are included in the main pipeline. These data augmented techniques are based on variations in color and stain, symmetric transformation and changes in brightness and contrast as shown in Fig. 5 and Fig. 6. The color, brightness, contrast variations and symmetric transformation are performed by using the fast and flexible image augmentation techniques (A. Buslaev and Kalinin, 2018). In order to increase the color variability, RGB (red,

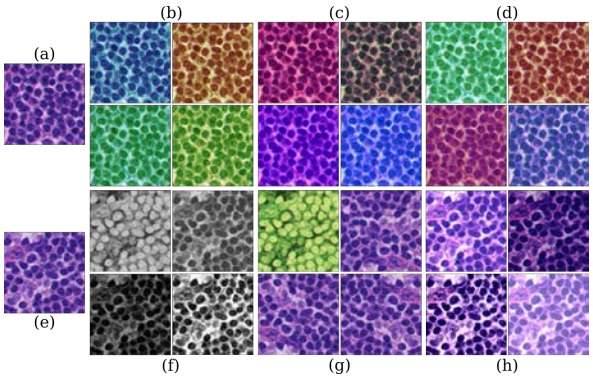


Figure 5: An example of data augmentation, on (a) and (e) original images, by (b) rbg channel shuffle, (c) rgb channel shifting, (d) hsv channel shifting, (f) brightness, contrast, inversion operations on gray version, (g) rgb inversion and symmetric operations and (h) brightness and contrast operations on rgb.

green, blue) and HSV (hue, saturation, value) channels of each image are shifted with randomly generated values. For RBG channels, shifting values are between [-80:80, -45:45, -40:40] whereas the HSV channels are randomly shifted with a ranges of [-180:180, -20:20, -27:27]. Different new colors of the each image are also produced by randomly shuffling the RGB channels. The brightness and contrast variations are produced with ranges between [-1.2:1.2] and [-0.9:0.9] respectively. The Contrast Limited Adaptive Histogram Equalization (CLAHE) and random gamma correction are also applied to variate brightness and contrast. For the symmetric transformation, images are randomly rotated between [-100:100] degrees and flipped horizontally and vertically. Similarly, images from each training batch are randomly inverted and converted to grayscale to produce more variability in the training batch. The gray version are also augmented with different symmetric transformation, brightness and contrast variations. The WSIs from Chamelyon17 are stained with H&E stains, the linear transformation of these stain to RGB space without the background can be represented by stain matrix S as given in Eq.1.

$$S = \begin{bmatrix} H_R & H_G & H_B \\ E_R & E_G & E_B \end{bmatrix} \tag{1}$$

Where the first and second rows are corresponding to the RGB compoents of the H&E stains respectively. The stain matrix S is estimated by Macenko et al. (2009) with the help of two largest singular values of decomposition vector. The RGB components of H&E stains are individually estimated by considering the light absorption coefficient C for each stain. Therefore, new staining concentration in each image is produced by variation of C coefficient by using Beer-Lambert law in Eq.2.

$$I_{RGB} = I_0 \, exp\,(-S_{RGB}.\,C) \tag{2}$$

In Fig. 6 an example of different variations produced by stain augmentation of the same image are shown.

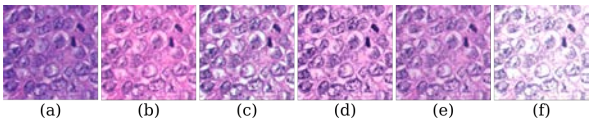


Figure 6: Stain augmentation, on (a) original image by obtaining different stain augmented versions from (b) to (g).

3.3. Stain Color Normalization

In order to reduce the color variations across data, the stain color normalization methods are used to homogenized train, validation, internal test and external test sets in each fold or sub data. The stain color distributions across different centers have been homogenized to a single target or template image. The stain normalization overcomes the color variance and model should perform well even on unseen stains due to the uniform stain color distribution. In this thesis, three stain color normalization methods are evaluated on tumor and normal tissue classification task (Byfield, 2019). Firstly, the histogram specification or matching is evaluated for stain color normalization in our data (Coltuc et al., 2006; Gonzalez and Woods, 2006). Where histogram of each patch in the data is matched to the histogram of specified target or template image with the help of cumulative distribution function as given in Eq.3 and Eq.4.

$$cdf_{src(R,G,B)}(s_i) = \sum_{j=0}^{i} p_{src(R,G,B)}(s_j) \tag{3}$$

$$cdf_{tmp(R,G,B)}(t_i) = \sum_{j=0}^{i} p_{tmp(R,G,B)}(t_j) \quad (4)$$

Where i is total number of gray level in each channel of RGB image, $cdf_{src(R,G,B)}(s_i)$ and $cdf_{tmp(R,G,B)}(t_i)$ are cumulative distribution functions of each gray level s_i and t_i in source and template image respectively. Similarly, $p_{src(R,G,B)}(s_j)$ and $p_{tmp(R,G,B)}(t_j)$ are representing the probability density function of each gray level s_i and t_i in source and template image respectively. Probability density function are calculated from the histogram of the both images, by considering the ratio of the frequency of the gray value to the total number of the pixels in the channel. Finally, t_i value in the template image is mapped to s_i value in each source image for a uniform stain color distribution. Similarly, stain color distribution among all the patches from train, validation, internal and external tests are specified to a single template image distribution in RGB channels.

Secondly, Macenko et al. (2009) stain color normalization approach is evaluated where the normalization is performed at H&E channels. All the patches from each fold of data are mapped to a template image by estimating stain colors in optical density. Where singular value decomposition (SVD) is used to get the optimal stain vectors from both input and template images to perform the linear per channel normalization based on the 99th percentile intensity values. Thirdly, Reinhard et al. (2001) stain color normalization is evaluated, the method is mainly based on color distribution model in each channel of the *Lab* color space. Both source and template images are converted from RGB to *Lab* color space. Then mean and standard deviation of each channel of both images are calculated. The color distribution of the template image is transferred to the source images in the *Lab* color space by using the calculated mean and standard deviation as shown in Eq.5.

$$I_{norm(L,a,b)} = \frac{[I_{src(L,a,b)} - \mu_{src(L,a,b)}] \times std_{tmp(L,a,b)}}{std_{src(L,a,b)} \times \mu_{tmp(L,a,b)}} \quad (5)$$

Where $I_{src(L,a,b)}$, $\mu_{src(L,a,b)}$ and $std_{src(L,a,b)}$ are the respective channels, mean and standard deviation values of source image in *Lab* color space. Similarly, $\mu_{tmp(L,a,b)}$ and $std_{tmp(L,a,b)}$ are representing mean and standard deviation of the corresponding template image respectively. Finally, the normalized image in *Lab* color space $I_{norm(L,a,b)}$ is converted back to RGB. In this thesis, it is also hypothesized that CNN models are efficient to learn morphological patterns in histopathology images and removing stain color information could improve the performance. Therefore, beside above mentioned stain color normalization methods, images are also converted to grayscale and grayscale histogram stretched versions to evaluate them on CNN model. The Fig. 7 presents few examples of the different patches from five different centers with color variability and then above mentioned normalization methods are used for uniform stain color distribution.

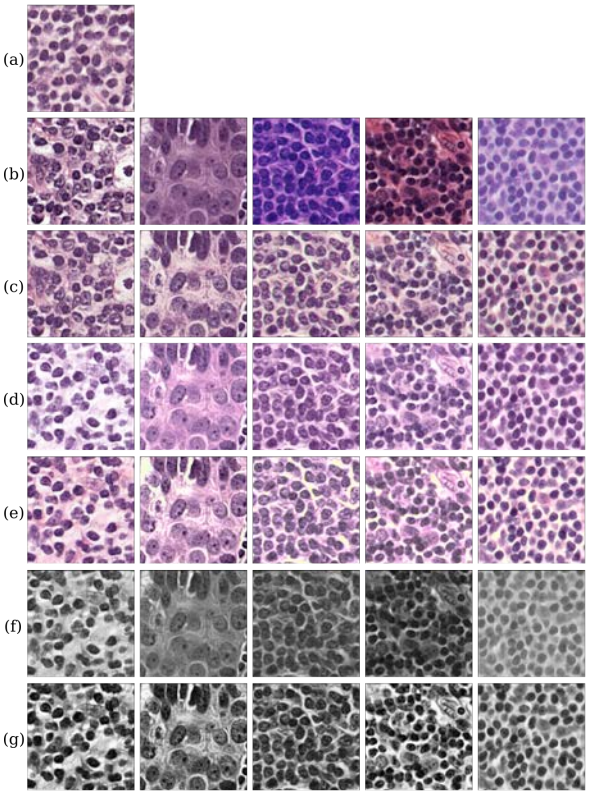


Figure 7: Stain color normalization, in the first row (a) A target or template image is used to distribute stain color homogeneously across (b) original images with the help of different color normalization methods ((c) Histogram specification, (d) Macenko, (e) Reinhard) and also original images were homogenized to (f) gray-scale and (g) gray-scale histogram stretched images.

3.4. Convolution Neural Network Classifier

In this thesis, MobileNetV2 architecture is used as a CNN classifier which belongs to the family of second generation computer vision networks (Sandler et al., 2018). Such networks are designed to perform detection, classification and segmentation related tasks at a very low computational costs with the aim to integrate them into personal mobile devices. The MobileNetV2 is the advance version of MobileNetV1 with more robustness and stability (Howard et al., 2017). The block diagram of the network is shown in Fig. 8, where the network contains the initial fully convolution layer with 32 filters, followed by 16 linear bottleneck blocks having a shortcut connections between them. These building blocks encode intermediate inputs and outputs with inner layer's ability to transform pixels to higher level descriptors with faster training and better accuracy. In contrast to conventional networks, full convolution operation is replaced with factorized depth-wise and point-wise convolution operations. These operations are performed with lightweight filtering and 1×1 convolution

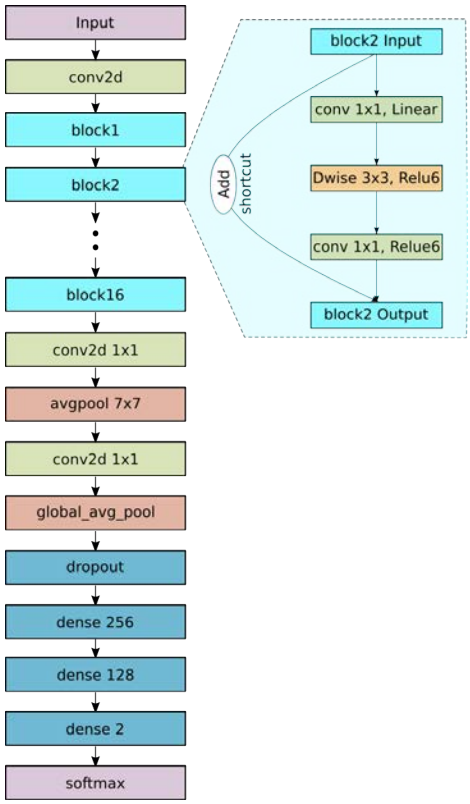


Figure 8: MobileNetV2, convolution neural network architecture.

operations to build new features through linear combinations of input channels. In the network Relu6 is used as non linearity due to its robustness when low precision computational tasks are required. The filtering operations throughout the network are performed by 3×3 standard kernel size. The network is designed with the expansion mechanism of the layers at each bottleneck block to improve the performance with small network size. The input channels to the layers are intermediately expended with a factor of 6 throughout the network. Output of the last bottleneck block is applied to two 1×1 convolution layers with an average pooling layer between them. In order to make the network useful for our class label classification task, the original network is extended by adding two dense layers of 256 and 128 neurons with Relu as an activation function. The network is extended upon adding globing average pooling and a dropout layer with 50% drop probability of neuron connections as shown in Fig. 8. Finally, a dense layer with 2 neurons with Softmax as an activation function is used to obtain the class predictions of the binary classification problem.

3.5. Domain Adversarial Network

In this thesis, it is hypothesized that training the CNN network to adopt the domain information while training for tumor and normal tissues classification task could improve the performance with less efforts as compared

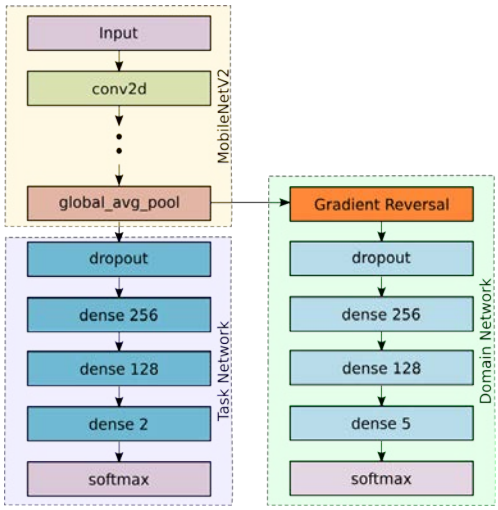


Figure 9: Domain adversarial convolution neural network.

to stain color normalization on an heterogeneous data. Therefore, inspired by Ganin et al. (2016), the CNN classifier is also trained to learn domain information while training for class label classification task. An extra domain network is added to the CNN classifier (as in Fig.8) by introducing the gradient reversal layer as shown in Fig. 9. The domain network is consisting of the two dense layer with the same configurations as task network. However, the last dense layer contains 5 neurons with Softmax as an activation function due the reason that we have 5 domains or histopathology centers responsible for stain heterogeneity across the data. While training the original network (MobileNetV2) is shared between task and domain networks acting as class label classifier and domain classifier. Both classifiers together form the domain adversarial network, where the network learns the features that do not consider the domain of the training samples. In the network, the gradient reversal layer does not contains any parameter to be updated, however, it remains unchanged during forward propagation and reverses the gradient by multiplying a negative scalar while backpropagation. In the domain adversarial training, same training configurations are used as described in section 3.6, the gradient loss is updated on each batch of the train data.

3.6. Training Phase

The pretrained MobileNetV2 on ImageNet is fine tuned on our prepared dataset of breast lymph node for tumor and normal tissue classes with a combination of stain color augmentation and normalization methods as well as domain adversarial experiments. The model is fine tuned over the patches from train set by preserving all the layers and its weights with the exception of last two fully connected layers in each case (task or domain adversarial). Then the training is performed by enabling all the layers of the network as trainable by

minimizing the cross-entropy loss with Stochastic gradient descent (SGD) optimizer. The initial learning rate of 1×10^{-3} is used and during the training it is decreased on every 5 epochs with a factor of 0.5. The model is trained with 16 samples per batch until 25 epochs with an early stop strategy. Upon training, the corresponding trained weights with highest validation accuracy is selected. The training for each experiment is repeated 5 times and averaged performance measures with standard deviation are recorded. The implementation of the network is performed in Keras plateform with Tensorflow on the back-end.

3.7. Performance and Statistical Measures

The outcome of the tumor and normal tissues classification problem is evaluated with two measures, AUC (area under the curve) and F1 score. AUC is the standard measure used in machine learning to validate the performance of a model for a certain task on a scale ranges between 0 to 1. AUC on a binary problem is calculated from the ROC (receiver operating characteristics) curve which is a visualization performance measure for a model. The ROC curve is a plot of true positive rate versus false positive rate of the classification probability with a threshold. Whereas the F1 score is the weighted average of the precision and recall. Precision is the ratio of the true positive to all positive predictions in the task and recall is the ratio of true positive to all predictions in the tumor class. Since, the F1 is considered as better performance measure over the accuracy in an uneven class problem, therefore, a slightly imbalance classes in our data set are the reason to evaluate the model on F1 score instead of accuracy. Both performance measures, AUC and F1 score are evaluated and reported on the internal and external test samples. Besides, paired McNemar's statistical test is applied to the class predictions of each preprocessing settings of the CNN classifier (McNemar, 1947; Raschka, 2018). The test is evaluated in order to obtain the most significant stain color normalization methods with or without augmentation on CNN classification when compare with the baseline (training of CNN classifier without any stain color normalization).

Table 3: Classification comparison on the PatchCamelyon dataset with our baseline with and without augmentation.

Method	Accuracy	AUC
B.Veeling et al.2018	0.898	0.963
Baseline	0.830	0.923
Baseline with augmentation	0.881	0.950

4. Results

In order to deal with stain color heterogeneity problem in the histopathology slides acquired from different centers, various experiments has been conducted. These experiments are the combinations of the stain color normalization augmentation methods.The proposed baseline CNN classifier is trained on the original images without applying any normalization method. However, the baseline is trained by using train samples with and without augmentation methods. Then each normalization method with and without augmentation is quantified with the baseline. Besides, the domain adversarial based training on the same data is performed by learning the domain information of the training samples. In the following sub-sections, the experimental results with each setting on five different sub data folds are described in detail.

4.1. Quantification of Data Augmentation

In this thesis, several suitable data augmentation techniques to histopathology images have been investigated as discussed in section 3.2. Before applying these data augmentation techniques to our prepared data, it is considered to quantify them on available CNN classification task to compare their performance with a benchmark. Therefore, the data augmentation techniques are evaluated on PatchCamelyon, a benchmark data set (Veeling et al., 2018). PatchCamelyon is the collection of patches from Camelyon16 pathology grand challenge (Veta et al., 2018), where the slides were prepared at two different pathology centers. In PatchCamelyon, 327680 patches of 96×96 pixels from tumor and normal tissues are equally distributed into train (262144), validation (32768) and test (32768) sets. A baseline on PatchCamelyon with proposed CNN classifier (as in Fig. 8) is developed with same parameters as described in 3.6 and results are presented with and without our chosen augmentation techniques. The performance comparison of our proposed baseline with (Veeling et al., 2018) are presented in Table 3. The performance is evaluated on AUC and accuracy that were originally used by Veeling et al. (2018). Both AUC and accuracy are improved from 0.923 to 0.950 and 0.832 to 0.881 respectively when the baseline is trained on proposed data augmentation. The performance of our baseline along with proposed data augmentation methods is almost comparable with Veeling et al. (2018). This comparison supported the application of these augmentation techniques to further experiments on the prepared data in order to enhance stain color variability while training CNN classifier to generalize well on external data.

4.2. Performance of stain color normalization methods

Experiments are conducted to train the proposed CNN classifier on selected stain color normalization methods as discussed in section 2.2. Stain color normalization based classification process is followed in all five sub-data folds that represent patches from five different pathology centers. The performance of the

Table 4: Experimental results of tumor and normal tissue classification on the internal and external test sets with different normalization, augmentation and domain adversarial using proposed CNN network. The values show the AUC and F1 scores averaged across 5 repetitions with standard deviation between the parenthesis along with p-values (paired McNemar's statistical test).

Data Fold	Normalization	Augmentation	Internal Test set		External Test set		p-value
			AUC	F1	AUC	F1	
1	Baseline	NA	0.847(0.016)	0.765(0.007)	0.860(0.006)	0.766(0.006)	-
	Histogram Specification	NA	**0.852(0.015)**	**0.773(0.018)**	0.862(0.014)	0.762(0.031)	<0.0001*
	Reinhard	NA	0.842(0.004)	0.751(0.016)	**0.871(0.007)**	**0.774(0.015)**	0.0206*
	Macenko	NA	0.841(0.005)	0.763(0.010)	0.827(0.021)	0.747(0.025)	0.0019*
	Grayscale	NA	0.834(0.020)	0.751(0.011)	0.845(0.011)	0.722(0.029)	0.0181*
	Grayscale-HS	NA	0.840(0.010)	0.773(0.012)	0.839(0.003)	0.765(0.003)	0.6870
	Domain Adversarial	NA	0.627(0.014)	0.607(0.016)	0.658(0.016)	0.565(0.020)	-
	Baseline	CS,BCS	0.867(0.009)	0.809(0.005)	0.874(0.003)	0.782(0.009)	-
	Histogram Specification	CS,BCS	0.857(0.005)	0.791(0.010)	0.862(0.018)	0.780(0.018)	0.0016*
	Reinhard	CS,BCS	0.869(0.003)	0.798(0.008)	**0.889(0.003)**	**0.810(0.001)**	0.0080*
	Macenko	CS,BCS	**0.874(0.002)**	**0.819(0.006)**	0.854(0.003)	0.787(0.005)	0.0308*
	Grayscale	BCS	0.855(0.005)	0.748(0.050)	0.842(0.014)	0.710(0.073)	<0.0001*
	Grayscale-HS	BCS	0.850(0.006)	0.777(0.008)	0.844(0.003)	0.772(0.003)	0.1098
	Domain Adversarial	BCS	0.530(0.018)	0.521(0.010)	0.551(0.043)	0.468(0.015)	-
2	Baseline	NA	**0.876(0.010)**	0.780(0.017)	0.816(0.009)	0.734(0.010)	-
	Histogram Specification	NA	0.867(0.017)	**0.785(0.018)**	**0.838(0.011)**	**0.748(0.006)**	<0.0001*
	Reinhard	NA	0.861(0.021)	0.775(0.010)	0.823(0.015)	0.746(0.005)	<0.0001*
	Macenko	NA	0.848(0.003)	0.759(0.003)	0.812(0.003)	0.734(0.008)	0.6863
	Grayscale	NA	0.860(0.003)	0.761(0.023)	0.814(0.005)	0.726(0.014)	0.2538
	Grayscale-HS	NA	0.855(0.003)	0.743(0.0292)	0.829(0.008)	0.692(0.030)	<0.0001*
	Domain Adversarial	NA	0.601(0.023)	0.575(0.012)	0.600(0.026)	0.571(0.025)	-
	Baseline	CS,BCS	0.876(0.005)	0.795(0.005)	0.834(0.006)	0.751(0.008)	-
	Histogram Specification	CS,BCS	0.866(0.004)	0.785(0.013)	**0.852(0.006)**	0.751(0.015)	0.1383
	Reinhard	CS,BCS	**0.888(0.006)**	**0.814(0.005)**	0.843(0.002)	**0.767(0.001)**	<0.0001*
	Macenko	CS,BCS	0.881(0.005)	0.781(0.021)	0.832(0.006)	0.727(0.006)	0.5349
	Grayscale	BCS	0.868(0.007)	0.776(0.006)	0.843(0.002)	0.761(0.009)	0.0236*
	Grayscale-HS	BCS	0.863(0.003)	0.772(0.014)	0.844(0.006)	0.755(0.016)	0.4895
	Domain Adversarial	BCS	0.582(0.036)	0.552(0.027)	0.610(0.020)	0.578(0.008)	-
3	Baseline	NA	**0.862(0.006)**	0.772(0.015)	0.797(0.013)	0.646(0.023)	-
	Histogram Specification	NA	0.850(0.002)	**0.805(0.003)**	0.858(0.002)	**0.773(0.008)**	<0.0001*
	Reinhard	NA	0.853(0.006)	0.777(0.012)	**0.860(0.005)**	0.745(0.031)	<0.0001*
	Macenko	NA	0.829(0.011)	0.748(0.025)	0.824(0.026)	0.735(0.010)	0.0048*
	Grayscale	NA	0.847(0.007)	0.760(0.021)	0.828(0.003)	0.731(0.017)	0.5986
	Grayscale-HS	NA	0.851(0.002)	0.784(0.012)	0.830(0.011)	0.733(0.013)	0.1465
	Domain Adversarial	NA	0.578(0.031)	0.551(0.022)	0.633(0.016)	0.594(0.014)	-
	Baseline	CS,BCS	0.866(0.003)	0.789(0.013)	0.830(0.011)	0.742(0.003)	-
	Histogram Specification	CS,BCS	0.863(0.001)	0.804(0.008)	**0.877(0.006)**	**0.798(0.003)**	0.0090*
	Reinhard	CS,BCS	**0.869(0.002)**	**0.812(0.004)**	0.872(0.004)	0.796(0.001)	0.0001*
	Macenko	CS,BCS	0.850(0.002)	0.799(0.002)	0.857(0.004)	0.787(0.007)	0.0390*
	Grayscale	BCS	0.857(0.014)	0.780(0.024)	0.827(0.014)	0.733(0.008)	0.3717
	Grayscale-HS	BCS	0.851(0.001)	0.797(0.008)	0.838(0.009)	0.757(0.012)	0.6867
	Domain Adversarial	BCS	0.461(0.028)	0.390(0.032)	0.633(0.061)	0.562(0.055)	-
4	Baseline	NA	0.839(0.010)	0.766(0.009)	0.839(0.028)	0.773(0.017)	-
	Histogram Specification	NA	0.835(0.008)	0.751(0.024)	**0.913(0.013)**	**0.831(0.011)**	0.0665
	Reinhard	NA	**0.845(0.011)**	**0.786(0.011)**	0.883(0.013)	0.807(0.006)	<0.0001*
	Macenko	NA	0.810(0.006)	0.717(0.043)	0.867(0.043)	0.782(0.033)	0.0048*
	Grayscale	NA	0.815(0.018)	0.736(0.048)	0.876(0.007)	0.759(0.045)	0.5986
	Grayscale-HS	NA	0.833(0.012)	0.739(0.020)	0.898(0.010)	0.806(0.016)	0.1465
	Domain Adversarial	NA	0.591(0.025)	0.548(0.023)	0.654(0.023)	0.590(0.032)	-
	Baseline	CS,BCS	0.835(0.003)	0.777(0.005)	0.857(0.005)	0.799(0.006)	-
	Histogram Specification	CS,BCS	0.829(0.015)	0.733(0.040)	0.890(0.024)	0.786(0.032)	<0.0001*
	Reinhard	CS,BCS	**0.848(0.001)**	**0.781(0.019)**	0.900(0.003)	0.832(0.003)	0.0154*
	Macenko	CS,BCS	0.834(0.001)	0.771(0.003)	0.870(0.001)	0.812(0.003)	0.1813
	Grayscale	BCS	0.827(0.009)	0.736(0.033)	0.897(0.016)	0.801(0.004)	<0.0001*
	Grayscale-HS	BCS	0.842(0.010)	0.781(0.009)	**0.902(0.034)**	**0.838(0.040)**	0.0998*
	Domain Adversarial	BCS	0.584(0.040)	0.531(0.007)	0.637(0.054)	0.581(0.050)	-
5	Baseline	NA	0.905(0.005)	0.819(0.002)	**0.852(0.014)**	0.736(0.052)	-
	Histogram Specification	NA	0.903(0.003)	0.828(0.007)	0.847(0.008)	**0.766(0.003)**	0.0098*
	Reinhard	NA	**0.920(0.006)**	**0.849(0.006)**	0.841(0.011)	0.752(0.005)	0.0341*
	Macenko	NA	0.885(0.014)	0.802(0.006)	0.830(0.003)	0.719(0.034)	0.0122*
	Grayscale	NA	0.887(0.004)	0.808(0.001)	0.728(0.043)	0.550(0.083)	0.1282
	Grayscale-HS	NA	0.897(0.005)	0.824(0.009)	0.811(0.009)	0.705(0.035)	0.2006
	Domain Adversarial	NA	0.836(0.006)	0.764(0.020)	0.777(0.006)	0.716(0.008)	-
	Baseline	CS,BCS	0.911(0.004)	0.841(0.007)	0.867(0.008)	0.786(0.005)	-
	Histogram Specification	CS,BCS	0.901(0.002)	0.830(0.003)	0.872(0.021)	0.794(0.021)	<0.0001*
	Reinhard	CS,BCS	**0.918(0.007)**	**0.844(0.005)**	**0.882(0.002)**	**0.801(0.007)**	0.0660*
	Macenko	CS,BCS	0.900(0.003)	0.823(0.003)	0.851(0.014)	0.775(0.011)	0.0291*
	Grayscale	BCS	0.892(0.004)	0.822(0.002)	0.804(0.005)	0.719(0.009)	<0.0001*
	Grayscale-HS	BCS	0.898(0.001)	0.838(0.003)	0.857(0.013)	0.766(0.013)	0.2979
	Domain Adversarial	BCS	0.771(0.025)	0.710(0.029)	0.763(0.005)	0.670(0.014)	-

Note: NA: No Augmentation, CS: Color and Stain, BCS: Brightness, Contrast and Symmetric, HS: Histogram Stretched
*p-value<0.05 (paired McNemar's Test)

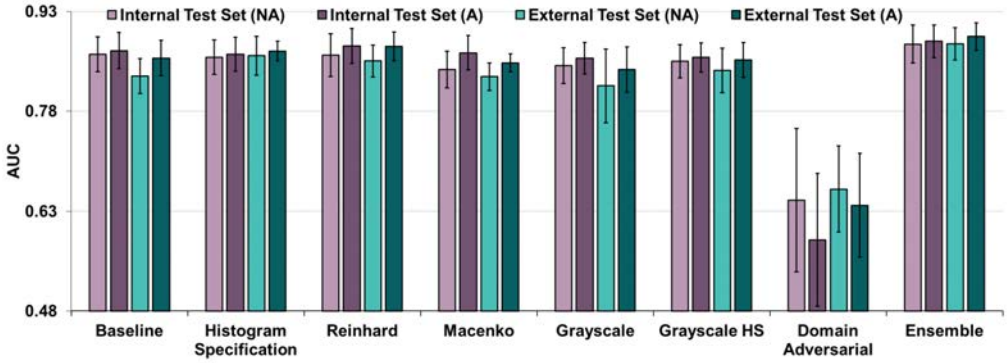


Figure 10: Average AUC measure and standard deviation across five data folds on proposed classification task to deal stain color heterogeneity by using normalization and augmentation methods. NA: No Augmentation, A: Augmentation

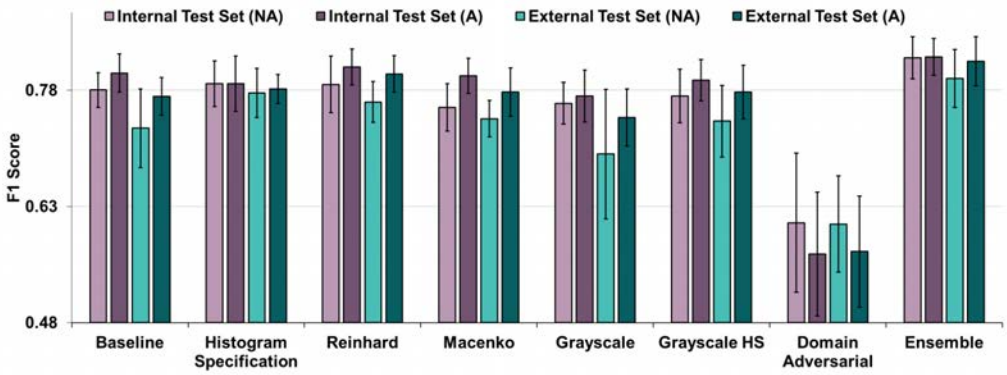


Figure 11: Average F1 score and standard deviation across five data folds on proposed classification task to deal stain color heterogeneity by using normalization and augmentation methods. NA: No Augmentation, A: Augmentation

each stain normalization method along with the baseline is presented in Table 4. The evaluation is based on averaged AUC and F1 score measures along standard deviation of 5 training repetitions with best validated model weights for internal and external test sets of each data fold. In order to make the results more interpretive, the average performance scores across all five data folds for each of the technique including baseline are presented in Fig.10 and Fig.11. Where the baseline obtained an average AUC across five data fold as 0.866 ± 0.026 and 0.833 ± 0.026 on internal and external test samples respectively. The corresponding F1 score is measured as 0.780 ± 0.022 and 0.731 ± 0.051. Among stain color normalization methods, none of the techniques scored higher then the baseline when evaluated on internal test set with AUC. However, the histogram specification and Reinhard outperformed on external test set with 0.864 ± 0.029, 0.856 ± 0.029 AUC and 0.777 ± 0.032, 0.765 ± 0.026 F1 scores respectively.

4.3. Significance of Data Augmentation Techniques

Upon evaluation of the proposed classifier on the normalization method to minimize the impact of stain color heterogeneity, the investigated data augmentation techniques are applied to the baseline as well as normalization methods. The detailed experimental results across each data fold are listed in Table 4 whereas the average score is presented in Fig.10 and Fig.11. It is evident from the results that all normalization techniques including baseline showed improved performance when evaluated on internal and external test samples with data augmentation. The average AUC across all data folds on the baseline is raised to 0.871 ± 0.027 and 0.860 ± 0.027 with corresponding F1 scores of 0.802 ± 0.025 and 0.772 ± 0.024 on internal and external test sets respectively. Reinhard normalization with augmentation outperformed over others including baseline with 0.878 ± 0.026, 0.877 ± 0.022 AUC and 0.810 ± 0.023, 0.801 ± 0.024 F1 score in both internal and external test data respectively. After Reinhard, the histogram specification with data augmentation achieved 0.871 ± 0.022 AUC and 0.782 ± 0.019 F1 score on external test set. However, Macenko showed better performance then histogram specification on internal test set, with 0.868 ± 0.026 AUC and 0.799 ± 0.023 F1 score.

4.4. Performance on Ensemble Predictions

The Mc Nemar's significance test is performed to assess the most significant preprocessing settings on the classification compare to baseline on the obtained results as shown in Table 4. The test is evaluated on the

combinations of augmentation and normalization methods with their prediction on both internal and external test sets. The calculated p-values of the each technique are presented Table 4 where last column contains the maximum p-value of internal and external test sets. From the statistical evaluation, histogram specification and Reinhard with and without augmentation on both test sets obtain an average p-value <0.017 at significance level of 0.05. Smaller p-value then significance level showed that model on both methods predicted better then others. Therefore, predictions of both histogram specification and Reinhard are ensemble by fusing their probabilities through element-wise multiplication rule as shown inf Fig.12 (Liu et al., 2019). Ensemble results on AUC and F1 measures are shown in Fig.10 and Fig.11. Average ensemble AUC scores without augmentation 0.881 ± 0.029, 0.882 ± 0.024 and with augmentation 0.885 ± 0.025, 0.893 ± 0.022 are obtained on internal and external tests respectively. The corresponding F1 scores are recorded as 0.822 ± 0.027, 0.795 ± 0.037 and 0.823 ± 0.024, 0.817 ± 0.032 respectively.

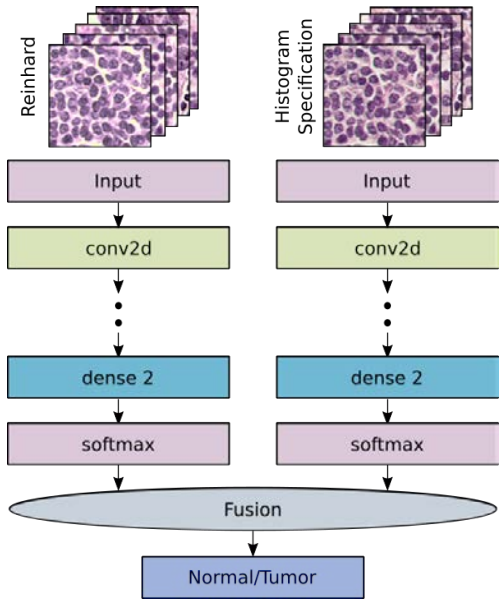


Figure 12: Fusion of CNN classifier trained on normalized images by Histogram Specification and Reinhard.

4.5. Performance of Domain Adversarial

In domain adversarial, the network is trained without any stain color normalization to train the network to learn domain information of the training samples. Applying the normalization to the data removes the stain color heterogeneity which is domain information that we want the domain adversarial should learn. However, non stain color augmentation methods such as symmetric transformations, brightness and contrast variations are included while training the domain adversarial network. Both results (with and without augmentation) on each data fold are listed in the Table 4 whereas the average performance is plotted in Fig.10 and Fig. 11. The overall performance of the domain adversarial is below 0.67 on both AUC and F1 score with the evaluation on internal and external test. However, better results obtained when the network is trained on original data compared to augmentation methods.

Table 5: Normalization time for each method on a randomly taken patch with a resolution of 224×224 pixels.

Normalization Technique	Time (Seconds)
Histogram Specification	0.031
Reinhard	0.198
Macenko	0.298
Grayscale	0.001
Grayscale-HS	0.002

Note: HS: Histogram Stretching

5. Discussion

In this thesis, the stain color heterogeneity in the histopatholgy images has been explored by minimizing its effects on a CNN based classification task with the help of various stain color normalization techniques, data augmentation methods and domain adversarial training. For a combination of stain color normalization and data augmentation methods, a CNN classifier is trained to classify tumor and normal tissues acquired from Camelyon17 grand challenge, containing stain color variability from five histology centers. The prepared dataset is distributed into five sub data folds by considering each time the images from a center as an external source of patches for an external test evaluation. A baseline classifier is trained on each fold of data without any stain normalization or augmentation technique. Then in the first series of the experiments, the patches from each fold are normalized in terms of stain color before passing to the CNN classifier to quantify their performance with the baseline. The stain color is normalized by histogram specification, Reinhard, Macnenko, grayscale and grayscale histogram stretched. By analyzing the obtained results on the normalization approaches, it is evident that the classifier performed well on external data when normalized with histogram specification and Reinhard. However, the Macenko and grayscale enhanced version shown almost identical performance on external test. Afterwords, in the second series of the experiments, the assessed data augmentation techniques are applied to the baseline and normalization based training. By introducing augmentation in training, we were able to improve the overall performance of the classifer, however, histogram specification and Reinhard again performed well on the external test set among other techniques. Then statistical test is performed to obtain top normalization methods

along with augmentation to ensemble their probabilities. Where the McNemar's paired test also validated both histogram specification and Reinhard with a significance difference from the baseline at class probability level. Therefore, in the third series of experiments, the probabilities of both outperformed methods along with augmentation are fused by element-wise multiplication. Interestingly, the CNN classifier learned different features on both normalization methods and results are improved on both internal and external test sets when their probabilities are fused. The best AUC and F1 score on external test are obtained as 0.893 ± 0.022 and 0.817 ± 0.032 respectively. During above experiments, it is observed that Macenko normalization method was computationally expensive as compared to other normalization methods (see Table 5). The possible reason was its lengthy calculations for stain vectors of both template and original image in the optical density space. It was also observed that in few cases where the size of the nuclei is larger then the normal the Macenko treated both nuclei and connective tissues similar in terms of staining which makes it difficult to differentiate them.
In the beginning of this study, it was hypothesized that CNN classifier could learn better on pattern instead of colors. In order to evaluate this hypothesis, the stain color was removed by converting the patches to grayscale versions. In few cases the grayscale versions have shown better performances, however, overall results in above experiments are evident that the stain colors are important and effects the CNN classification based decisions along with morphological patterns. In addition to experiments on stain color normalization and augmentation methods, the domain adversarial network was also trained to learn the domain information along the class label classification. The ultimate goal of the domain adversarial based experiments was to minimize the effects of domain when training the CNN classifier for the class label classification task. However, domain adversarial could not converge well on our data set. It was observed in the data that heterogeneity does not exits across the centers only but the stain color is also heterogeneous within the slides belonging to a center. Intra-slide and inter-slide heterogeneities were found in some centers which makes it difficult for the domain adversarial network to stay at a decision while converging. The intra-slide heterogeneity could exist due to the temperature or stains used during slide preparation. Whereas the inter-slide heterogeneity arising from difference in thickness and amount of stain absorbed by the certain areas within a slide. In overall picture, this could lead to the generation of the several heterogeneous domains with the images originating from single (domain) center. Therefore, in such heterogeneous conditions, the stain color normalization along with data augmentation are useful to produce better performance on the data from external source which is also evident from the first series of experiments.

6. Conclusions

The findings of this work can be be summarized into four folds. Firstly, the investigation of the publicly available histopathology databases of various tissues. Secondly, the suitable collection from Camelyon17 is used to quantify various stain color normalization and augmentation methods. The quantification is performed on a convolution neural network based binary classification task of normal versus tumor tissues on the extracted patches from whole slide images of the Camelyon17. Thirdly, combining best performed stain color normalization and augmentation method to improve the performance on the data set from a completely external source then samples used in train, validation and internal test. Following the above pipeline, five sub data folds are created and extensive experimentation on various stain color normalization along with augmentation methods are conducted. The experimental results shown significance performance of the histogram specification and Reinhard with augmentation on external test samples over Macenko and grayscale versions. The best results are obtained by ensemble probabilities of both methods. Lastly, along with class label classification task, a domain adversarial network is trained to learn domain information of the corresponding training samples. The network is evaluated on training with and without data augmentation methods. The domain adversarial could not achieve even the baseline performance due to inter and intra slide heterogeneity with a center, however, the method can be evaluated further by applying some future work suggestions.

7. Future Work

In present work, normalization is limited to a single template image, however, this work can be extended to several template images to analyze the robustness of the normalization method as well as by including more normalization techniques to the comparison. An other possible extension can be the application of the best evaluated methods to other tissues with heterogeneity such as colorectal and prostate for different diagnostics tasks. The performance of domain adversarial network can be evaluated on the same data by applying the best normalization method at the slide or center level. The domain adversarial based learning can also be evaluated on grayscale versions in order to learn morphological features in the gray domain.

8. Acknowledgments

Amjad Khan holds an Erasmus Mundus master's scholarship. The authors are thankful to Dr. Manfredo Atzori, Dr. Vincent Andrearczyk, Sebastian Otalora Montenegro, Mara Graziani and Dr. Oscar Jimenez for useful discussion during this work at HES-SO.

References

A. Buslaev, A. Parinov, E.K.V.I.I., Kalinin, A.A., 2018. Albumentations: fast and flexible image augmentations. ArXiv e-prints .

Alturkistani, H.A., Tashkandi, F.M., Mohammedsaleh, Z.M., 2015. Histological Stains: A Literature Review and Case Study. Global Journal of Health Science 8, 72. doi:10.5539/gjhs.v8n3p72.

Arvaniti, E., Fricker, K., Moret, M., Rupp, N., Fankhauser, C., Hermanns, T., Wey, N., Wild, P., Rüschoff, J., Claassen, M., 2018. Automated Gleason grading of prostate cancer via deep learning. European Urology Supplements 17, e3020–e3021. doi:10.1016/s1569-9056(18)33852-1.

Awan, R., Sirinukunwattana, K., Epstein, D., Jefferyes, S., Qidwai, U., Aftab, Z., Mujeeb, I., Snead, D., Rajpoot, N., 2017. Glandular Morphometrics for Objective Grading of Colorectal Adenocarcinoma Histology Images. Scientific Reports 7, 16852. doi:10.1038/s41598-017-16516-w.

Baldassarre, F., González Morín, D., Rodés-Guirao, L., . Deep Koalarization: Image Colorization using CNNs and Inception-Resnet-v2. Technical Report.

Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Ehteshami Bejnordi, B., Lee, B., Paeng, K., Zhong, A., Li, Q., Zanjani, F.G., Zinger, S., Fukuta, K., Komura, D., Ovtcharov, V., Cheng, S., Zeng, S., Thagaard, J., Dahl, A.B., Lin, H., Chen, H., Jacobsson, L., Hedlund, M., Cetin, M., Halici, E., Jackson, H., Chen, R., Both, F., Franke, J., Kusters-Vandevelde, H., Vreuls, W., Bult, P., van Ginneken, B., van der Laak, J., Litjens, G., 2019. From Detection of Individual Metastases to Classification of Lymph Node Status at the Patient Level: The CAMELYON17 Challenge. IEEE Transactions on Medical Imaging 38, 550–560. doi:10.1109/TMI.2018.2867350.

Bentaieb, A., Hamarneh, G., 2018. Adversarial stain transfer for histopathology image analysis. IEEE transactions on medical imaging 37, 792–802.

BenTaieb, A., Li-Chang, H., Huntsman, D., Hamarneh, G., 2017. A structured latent model for ovarian carcinoma subtyping from histopathology slides. Medical Image Analysis 39, 194–205. doi:10.1016/j.media.2017.04.008.

Bug, D., Schneider, S., Grote, A., Oswald, E., Feuerhake, F., Schüler, J., Merhof, D., 2017. Context-Based Normalization of Histological Stains Using Deep Convolutional Features. Springer International Publishing, Cham.

Byfield, P., 2019. Staintools. https://github.com/Peter554/StainTools.

Cho, H., Lim, S., Choi, G., Min, H., 2017. Neural Stain-Style Transfer Learning using GAN for Histopathological Images 80, 1–10.

Coltuc, D., Bolon, P., Chassery, J.M., 2006. Exact histogram specification. IEEE Transactions on Image Processing 15, 1143–1152. doi:10.1109/TIP.2005.864170.

Drelie Gelasca, E., Byun, J., Obara, B., Manjunath, B., 2008. Evaluation and benchmark for biological image segmentation, in: 2008 15th IEEE International Conference on Image Processing, IEEE. pp. 1816–1819. doi:10.1109/ICIP.2008.4712130.

Ehteshami Bejnordi, B., Litjens, G., Timofeeva, N., Otte-Holler, I., Homeyer, A., Karssemeijer, N., van der Laak, J.A., 2016. Stain Specific Standardization of Whole-Slide Histopathological Images. IEEE Transactions on Medical Imaging 35, 404–415. doi:10.1109/TMI.2015.2476509.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2016. Domain-adversarial training of neural networks. J. Mach. Learn. Res. 17, 2096–2030.

Gonzalez, R.C., Woods, R.E., 2006. Digital Image Processing (3rd Edition). Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. Neural Computation 9, 1735–1780. doi:10.1162/neco.1997.9.8.1735.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications .

Ioffe, S., Szegedy, C., 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, in: Proceedings of the 32nd International Conference on International Conference on Machine Learning, JMLR.org. pp. 448—-456.

Janowczyk, A., Basavanhally, A., Madabhushi, A., 2017. Stain normalization using sparse autoencoders (stanosa): Application to digital pathology. Computerized Medical Imaging and Graphics 57, 50–61.

Kather, J.N., Weis, C.A., Bianconi, F., Melchers, S.M., Schad, L.R., Gaiser, T., Marx, A., Zöllner, F.G., 2016. Multi-class texture analysis in colorectal cancer histology. Scientific Reports 6, 27988. doi:10.1038/srep27988.

Khan, A.M., Rajpoot, N., Treanor, D., Magee, D., 2014. A Nonlinear Mapping Approach to Stain Normalization in Digital Histopathology Images Using Image-Specific Color Deconvolution. IEEE Transactions on Biomedical Engineering 61, 1729–1738. doi:10.1109/TBME.2014.2303294.

Khosravi, P., Kazemi, E., Imielinski, M., Elemento, O., Hajirasouliha, I., 2018. Deep Convolutional Neural Networks Enable Discrimination of Heterogeneous Digital Pathology Images. EBioMedicine 27, 317–328. doi:10.1016/j.ebiom.2017.12.026.

Li, Z., Hu, Z., Xu, J., Tan, T., Chen, H., Duan, Z., Liu, P., Tang, J., Cai, G., Ouyang, Q., Tang, Y., Litjens, G., Li, Q., 2018. Computer-aided diagnosis of lung carcinoma using deep learning - a pilot study .

Litjens, G., Bandi, P., Ehteshami Bejnordi, B., Geessink, O., Balkenhol, M., Bult, P., Halilovic, A., Hermsen, M., van de Loo, R., Vogels, R., Manson, Q.F., Stathonikos, N., Baidoshvili, A., van Diest, P., Wauters, C., van Dijk, M., van der Laak, J., 2018. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. GigaScience 7. doi:10.1093/gigascience/giy065.

Liu, Y., Zhang, C., Cheng, J., Chen, X., Wang, Z.J., 2019. A multi-scale data fusion framework for bone age assessment with convolutional neural networks. Computers in Biology and Medicine 108, 161–173. doi:10.1016/J.COMPBIOMED.2019.03.015.

Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E., 2009. A method for normalizing histology slides for quantitative analysis, in: Biomedical Imaging: From Nano to Macro, 2009. ISBI'09. IEEE International Symposium on, IEEE. pp. 1107–1110.

Marinelli, R.J., Montgomery, K., Liu, C.L., Shah, N.H., Prapong, W., Nitzberg, M., Zachariah, Z.K., Sherlock, G.J., Natkunam, Y., West, R.B., Van de Rijn, M., Brown, P.O., Ball, C.A., 2008. The stanford tissue microarray database. Nucleic Acids Research 36, D871–7. doi:10.1093/nar/gkm861.

McNemar, Q., 1947. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12, 153–157. doi:10.1007/BF02295996.

NIH, 2007. Understanding Cancer, National Institutes of Health (US).

Pantanowitz, L., Evans, A., Pfeifer, J., Collins, L., Valenstein, P., Kaplan, K., Wilbur, D., Colgan, T., 2011. Review of the current state of whole slide imaging in pathology. Journal of Pathology Informatics 2, 36. doi:10.4103/2153-3539.83746.

Raschka, S., 2018. Mlxtend: Providing machine learning and data science utilities and extensions to pythons scientific computing stack. The Journal of Open Source Software 3. doi:10.21105/joss.00638.

Reinhard, E., Ashikhmin, M., Gooch, B., Shirley, P., 2001. Color transfer between images. IEEE Computer Graphics and Applications 21, 34–41. doi:10.1109/38.946629.

Roux, L., Racoceanu, D., Capron, F., Calvo, J., Attieh, E., Naour, G.L., Gloaguen, A., 2014. Mitos & atypia. Technical Report. Paris.

Roy, S., kumar Jain, A., Lal, S., Kini, J., 2018. A study about color normalization methods for histopathology images. Micron 114, 42–61. doi:10.1016/j.micron.2018.07.005.

Samsi, S., Jones, M., Kepner, J., Reuther, A., 2018. Colorization of H&E stained tissue using Deep Learning, in: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE. pp. 640–643. doi:10.1109/EMBC.2018.8512419.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.,

2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE. pp. 4510–4520. doi:10.1109/CVPR.2018.00474.

Shaban, M.T., Baur, C., Navab, N., Albarqouni, S., 2018. StainGAN: Stain Style Transfer for Digital Histological Images , 1–8.

Shamir, L., Orlov, N., Mark Eckley, D., Macura, T.J., Goldberg, I.G., 2008. IICBU 2008: A proposed benchmark suite for biological image analysis. Medical and Biological Engineering and Computing 46, 943–947. doi:10.1007/s11517-008-0380-5.

Sirinukunwattana, K., Pluim, J.P.W., Chen, H., Qi, X., Heng, P.A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U., Böhm, A., Ronneberger, O., Cheikh, B.B., Racoceanu, D., Kainz, P., Pfeiffer, M., Urschler, M., Snead, D.R.J., Rajpoot, N.M., 2016. Gland Segmentation in Colon Histology Images: The GlaS Challenge Contest .

Smith, P., Reid, D.B., Environment, C., Palo, L., Alto, P., Smith, P.L., 1979. A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems, Man, and Cybernetics 9, 62–66. doi:10.1109/TSMC.1979.4310076.

Tam, A., Barker, J., Rubin, D., 2016. A method for normalizing pathology images to improve feature extraction for quantitative pathology. Medical Physics 43, 528–537. doi:10.1118/1.4939130.

Tellez, D., Balkenhol, M., Karssemeijer, N., Litjens, G., van der Laak, J., Ciompi, F., 2018. H and e stain augmentation improves generalization of convolutional networks for histopathological mitosis detection, in: Medical Imaging 2018: Digital Pathology, International Society for Optics and Photonics. p. 105810Z.

Tellez, D., Litjens, G., Bandi, P., Bulten, W., Bokhorst, J.M., Ciompi, F., van der Laak, J., 2019. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology .

Vahadane, A., Peng, T., Sethi, A., Albarqouni, S., Wang, L., Baust, M., Steiger, K., Schlitter, A.M., Esposito, I., Navab, N., 2016. Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images. IEEE Transactions on Medical Imaging 35, 1962–1971. doi:10.1109/TMI.2016.2529665.

Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M., 2018. Rotation equivariant CNNs for digital pathology .

Veta, M., Heng, Y.J., Stathonikos, N., Bejnordi, B.E., Beca, F., Wollmann, T., Rohr, K., Shah, M.A., Wang, D., Rousson, M., Hedlund, M., Tellez, D., Ciompi, F., Zerhouni, E., Lanyi, D., Viana, M., Kovalev, V., Liauchuk, V., Phoulady, H.A., Qaiser, T., Graham, S., Rajpoot, N., Sjöblom, E., Molin, J., Paeng, K., Hwang, S., Park, S., Jia, Z., Chang, E.I.C., Xu, Y., Beck, A.H., van Diest, P.J., Pluim, J.P.W., 2018. Predicting breast tumor proliferation from whole-slide images: the TUPAC16 challenge doi:10.1016/j.media.2019.02.012.

WHO, 2018. Latest global cancer data .

Xu, Y., Jia, Z., Wang, L.B., Ai, Y., Zhang, F., Lai, M., Chang, E.I.C., 2017. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. BMC Bioinformatics 18, 281. doi:10.1186/s12859-017-1685-x.

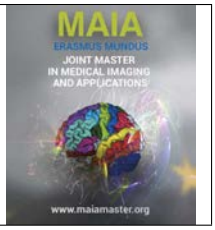
Yagi, Y., 2011. Color standardization and optimization in Whole Slide Imaging. Diagnostic Pathology 6, S15. doi:10.1186/1746-1596-6-S1-S15.

Zheng, Y., Jiang, Z., Zhang, H., Xie, F., Shi, J., Xue, C., 2019. Adaptive color deconvolution for histological WSI normalization. Computer Methods and Programs in Biomedicine 170, 107–120. doi:10.1016/j.cmpb.2019.01.008.

Zhong, Q., Guo, T., Rechsteiner, M., Rüschoff, J.H., Rupp, N., Fankhauser, C., Saba, K., Mortezavi, A., Poyet, C., Hermanns, T., Zhu, Y., Moch, H., Aebersold, R., Wild, P.J., 2017. A curated collection of tissue microarray images and clinical outcome data of prostate cancer patients. Scientific Data 4, 170014. doi:10.1038/sdata.2017.14.

Medical Imaging and Applications

Master Thesis, June 2019



Brain Age Prediction from MRI and MEG Data

Oleh Kozynets, Alexandre Gramfort, Denis Engemann

INRIA Saclay-Ile de France, Bat. Turing, 1 Rue Honor d'Estienne d'Orves, 91120 Palaiseau

Abstract

The difference between the chronological age of a person and the age predicted based on structural and functional magnetic resonance imaging (MRI) data was proved to indicate for neurological and psychiatric diseases. The main objective of the present investigation was to determine whether the combination of MEG, structural and functional MRI could improve the brain age prediction accuracy. We achieved mean absolute error of 4.75 years and the coefficient of determination of $R^2 = 0.89$. The prediction model was built by stacking of the ridge regression models using the random forest algorithm. We evaluated the model performance on the largest available data repository combining multi-modal data (MEG, structural and functional MRI) of approximately 700 participants aged 18-87 years, i.e. the Cambridge Center for Aging and Neuroscience (Cam-CAN) data set.

Keywords: Age Prediction, MEG, MRI, fMRI, Machine Learning

1. Introduction

The World Population Ageing 2017 report prepared by the Department of Economic and Social Affairs of the United Nations Secretariat states that the number of people aged 60 years and over has considerably increased across the world and growth is projected to accelerate in the future (DoEaSA-P, 2017). The age-associated diseases challenge the established health care systems (Vos et al., 2012). Still, there are numerous challenges in understanding the complex biological processes underlying the molecular mechanisms of aging (López-Otín et al., 2013). Moreover, there is no formal notion of what constitutes the normal aging of the human brain (Shafto et al., 2014). Hence, the chronological age of a person does not provide a sufficient amount of information about the changes in the brain structure and functioning.

The brain, as the whole human body, goes through a series of changes intertwined with aging. In the work of (Pfefferbaum et al., 1994) structural magnetic resonance imaging (MRI) was used to identify the effects of aging on the over-all state of the human brain. The authors discovered that gray matter (GM) volume was increasing from birth until the age of 4, then it was gradually decreasing in the following decades. At the same time, white matter (WM) volume grew steadily until the age of 20, and remained constant during the rest of the lifespan. Cerebrospinal fluid (CSF) and ventricular volumes remained constant for the first 20 years of life, and increased with age. These development patterns reflect different neuronal processes, such as cell growth, myelination (i.e. the process of generating myelin), cell death, and cerebral atrophy (i.e. a loss of neurons and the connections between them). A more recent study, conducted by Good et al. (2001), arrived at similar conclusions, while reporting accelerated grey matter decline in some brain areas and significant microstructural changes in white matter in general. Storsve et al. (2014) discovered that the primary contributor to cortical volume reductions in aging was cortical thinning, while cortical surface area experienced less profound changes.

Neurodegenerative diseases, notably Alzheimers disease (AD), lead to subtle changes in neuroanatomical shape, complexity, and tissue characteristics (Ashburner et al., 2003). A study, conducted by Davatzikos et al. (2009), reported that patients with Alzheimers disease or mild cognitive impairment (MCI) showed accelerated brain atrophy relatively to cognitively normal individuals. Schizophrenia results in significant decrease in gray matter concentration in multiple cortical and subcortical regions (Meda et al., 2008). The pathological changes

of the brain, caused by numerous brain disorders, can be considered as an accelerated aging process, implying accelerated brain atrophy (Franke et al., 2010).

Another acclaimed method of brain investigation is functional magnetic resonance imaging (fMRI). It is a magnetic resonance imaging technique that estimates brain activity by measuring changes in the local level of blood oxygenation, which reflects the intensity of local brain activity (Poldrack et al., 2011). The study of the individual neuronal connections using rest-state fMRI (i.e. recorded while no explicit task is being performed) is emerging as a mainstream tool for identification of neurological and psychiatric illness (Castellanos et al., 2013). For example, Dennis and Thompson (2014) discovered that healthy aging and Alzheimers disease are associated with changes in functional connectivity (i.e. an assessment of the integration of brain activity across distant brain regions). Normal aging led to non-uniform decline in functional connectivity, while the decline due to Alzheimers disease was more profound.

Magnetoencephalography (MEG) and electroencephalography (EEG) are methods that allow to capture electrical activity within the brain with a high temporal and relatively good spatial resolution (Hansen et al., 2010). Therefore, these techniques provide another mean of investigation of the brain activity. For example, Hipp et al. (2012) observed highly structured brain-wide correlation of rest-state electrophysiological signals. M/EEG allow to directly record the electromagnetic signals related to the activity of neurons, while fMRI only captures hemodynamic response associated with the brain functioning. Additionally, Hipp and Siegel (2015) found no generic, brain-wide transfer function from hemodynamic correlation to the correlation of frequency-specific neuronal activity. Recent MEG and EEG studies of brain oscillatory activity showed correlation between aging and electrophysiological activity, e.g. the latency of brain responses measured by MEG was linked to age (Price et al., 2017). In the work of (Vlahou et al., 2014), it was shown that healthy aging is accompanied by a marked and linear decrease of resting-state activity in the slow frequency range (0.5–6.5 Hz).

There are numerous aspects that constitute the process of healthy brain aging. Regarding the wide acceptance of machine learning methods in neuroscience (Bzdok, 2017) and psychology (Yarkoni and Westfall, 2017), these techniques can be used to aggregate information about the aging process of the brain. The difference between a biological age (the hypothetical age of an organism, defined by measuring some aspect of the organisms biology) and the organisms chronological age can indicate of residual lifespan, functional capacity, and age-associated risks (Cole and Franke, 2017). Generally, a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention is referred to as a biomarker (Atkinson A.J. et al., 2001). Thus, the difference between the chronological age of a person and their age measured using imaging data can be utilized as a biomarker of neurological syndromes that emerge late in the lifespan (Liem et al., 2017). To realize it one requires a robust method of age prediction for healthy population. MEG, structural and functional MRI capture information about the brain anatomy and activity from different perspectives. Hence, in our study we want to investigate whether the brain age prediction accuracy may be increased by incorporating these sources of information together.

2. State of the art

2.1. Multimodal imaging data for age prediction

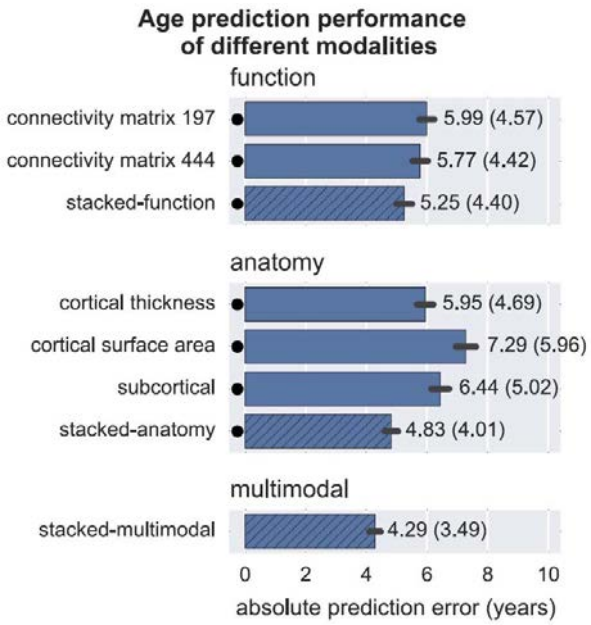


Figure 1: Age prediction of different modalities. The accuracy of every model was assessed using mean absolute error (MAE), standard deviation (STD) of age prediction was obtained as well. The figure is reproduced from Liem et al. (2017)

The number of research papers demonstrating that age prediction on MRI data using machine learning methods has both clinical and scientific relevance grows steadily (Cole and Franke, 2017). For example, in the work of Franke et al. (2010) the T1-weighted MRI images and a relevance vector machine (Tipping, 2001) were used to predict brain age. The authors achieved a mean absolute error of 5 years, while testing this approach on data collected from four different scanners for 650 healthy subjects, aged 19–86 years. Moreover, applying the proposed framework to people with mild AD gave a mean difference +10 years between the chronological and predicted age (Franke et al., 2010). In the later study, conducted by Franke and Gaser (2012) the

data of about 400 elderly subjects were analyzed. They discovered accelerated brain atrophy (+6 years difference with the chronological age) in patients with AD and the ones who had converted to AD within 3 years, while during follow-up an additional increase in atrophy resulted into a difference of about +9 years. According to Franke and Gaser (2012), the accumulated difference between the chronological and predicted age was due to disease severity and prospective cognitive decline. In the work conducted by Cole et al. (2015), it was discovered that traumatic brain injury (TBI) accelerates the rate of brain atrophy. A regression model was trained on structural MRI of 1,537 healthy individuals, and tested on 99 TBI patients. The difference between the predicted and chronological age for affected brains was 4.66 years for GM-based features and 5.97 years for WM-derived features. Moreover, in the investigation of Koutsouleris et al. (2014), the brains affected by the mental disorders were found to be 'older' than healthy brains. In detail, the predicted age was higher by 5.5 years in schizophrenia group (141 patients), followed by major depression (4.0 years, 104 patients), borderline personality disorder (3.1 years, 57 patients), and the individuals in at-risk mental states for psychosis (1.7 years, 89 patients). A machine learning model was trained on structural MRI scans of 800 healthy individuals using ν-support vector regression.

In the paper of Dosenbach et al. (2010), functional connectivity MRI studies were used to predict individuals brain maturity across development. Support vector regression machines (Drucker et al., 1997) were trained on 5 minutes long resting-state fMRI of 238 participants, aged from 7 to 30 years. The obtained functional maturation curve accounted for 55% of the sample variance.

As we can see, both structural and functional MRI data convey meaningful information about age. This motivated Liem et al. (2017) to combine these modalities. They used MRI data set obtained as part of LIFE-Adult-Study of the Leipzig Research Center for Civilization Diseases (LIFE) (Loeffler et al., 2015) to estimate brain age, as well as to find a relationship between predicted age and mental disorders. Effectively, data from 2354 individuals between 19 and 82 years old (1120 females and 1234 males) were processed. The robustness of the investigation results were additionally tested on the enhanced Nathan Kline Institute-Rockland Sample (NKI-RS) data set (Nooner et al., 2012).

The feature engineering was conducted separately for functional and structural MRI data. The fMRI predictors were obtained using the Nilearn package (Abraham et al., 2014). In detail, mean timeseries were obtained from cortical and subcortical regions of the BASC parcellation atlas (Bellec et al., 2010), functional connectivity between region pairs was calculated via Pearson correlation and further processed with Fisher's r-to-z transformation (Fisher, 1921). Connectivity matrices from 197 and 444 regions were used. In the case of structural MRI, *cortical thickness*, *cortical surface*, and *subcortical volumes* were estimated using the FreeSurfer software (Fischl, 2012). Available data were resampled into the fsaverage4 standard space, the data for the two hemispheres were concatenated. Therefore, age prediction models were trained on two functional maps of neural connections in the brain, namely *connectivity matrix 197* and *connectivity matrix 444*, derived from the functional MRI data; and three vectors of anatomical information originating from structural MRI.

In the work of Liem et al. (2017), a two-level architecture was proposed, for which the outputs from the models, trained using one of the previously obtained predictors, were aggregated into the final prediction by non-linear stacking. The single-source models were built using linear support vector regression (SVR). On the stacking level, random forest (RF) (Breiman, 2001) regression models were used to combine outputs from the linear models. In total three versions of multi-source models were investigated: *stacked-function* which merged predictions from the fMRI data, *stacked-anatomy* which merged predictions from the anatomical data, and *stacked-multimodal* which merged predictions from all available data. Mean absolute error (MAE) of the age predictions were incorporated for performance estimation, corresponding data for the LIFE data set can be found in figure 1. The coefficient of determination R^2 was estimated for every model as well.

The authors state that all models deal quite good with the task of age prediction, e.g. MAE = 5.25 years and $R^2 = 0.8$ for stacked-function model, and MAE = 4.83 years and $R^2 = 0.83$ for stacked-anatomy. They also note that the stacked models perform better than single source models and the combination of the structural and functional information achieves the best accuracy regarding the others. The mean absolute prediction error for the latter was equal to 4.29 years and $R^2 = 0.87$. Moreover, in the paper it was shown that the difference between the predicted age and chronological age correlates with cognitive impairment. At the same time, the authors discovered that the models do not generalize well to the data from a new site: the age prediction accuracy of the models trained on the LIFE samples deteriorates when tested on the NKI data. Though prediction performance tend to increase while training on the data from both sites, the same accuracy as in simulations with the LIFE data (Fig. 1) was not achieved (Liem et al., 2017).

2.2. *Electrophysiology data for age prediction*

Correlation between aging and electrophysiological activity was discovered in several recent studies (Price et al., 2017; Vlahou et al., 2014). This idea was further investigated by Khan et al. (2018). The authors used two data sets of resting-state MEG scans. The primary

data were acquired at Massachusetts general hospital (MGH), and consisted of 131 healthy subjects, between 7 and 29 years old. The second data set was intended for validation purposes only. It consisted of the data acquired from 31 young adults, between 21 and 28 years old, as part of the Open MEG Archive (OMEGA) by the McConnell Brain Imaging Centre of the Montreal Neurological Institute, and the University of Montreal (Niso et al., 2016). For all the available data MEG-specific processing techniques were applied. In detail, the authors performed noise suppression and motion correction using the signal space separation (SSS) method for the first data set, after that all data were mapped onto cortical space and obtained timeseries were averaged across labels. They were band-pass filtered and Hilbert transform was further performed on them. The frequency bands were set as: delta (1–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), beta (13–30 Hz), and gamma (31–80 Hz).

Synchronicity between different cortical labels were evaluated using orthogonal envelope correlation metric (Hipp et al., 2012) using an overlapping sliding window of 30 s with a stride of 1/8 of the window side (Khan et al., 2018) for each frequency band. The final result of the processing pipeline was a connectivity tensor of $n \times n \times n_{time} \times n_{bands}$ elements, where n is the total number of nodes, n_{time} is the number of sliding windows, and n_{bands} is the number of frequency bands. Signal to noise ratio (SNR) was increased by estimating the median of correlations across time-dimension. The adjacency matrix A, defined over a graph G such that any element A_{ij} is 1 if the edge e_{ij} between two nodes $v_i, v_j \in G$ exists and 0 otherwise; was obtained by thresholding and binarizing the connectivity matrix for each frequency band. The weighted adjacency matrix which preserves the correlation values above the same threshold was also computed.

Several metrics originating from graph theory were used to evaluate cortical networks: the average shortest path length, the degree and the local clustering coefficient for every network node, the average global and local efficiencies of information transfer in graph G, small world property (which measures the balance of short and long-range connections), betweenness centrality (which equals to the number of the shortest paths that pass through the vertex), resilience (which estimates the robustness of the network to the removal of the most heavily connected nodes).

The features that were related to age (according to Spearman correlation) were selected and combined into three sets. In detail, local network efficiency and small world property were found to increase with age in the beta band, while for the same band resilience of the network represented by connectivity matrix A was found to decrease with age. For the connectivity data in the gamma band, global network efficiency and resilience were increasing significantly with age, while small world property was significantly weakening. The predicted ages for all subjects from the regression were converted to the maturation indices using a scaling scheme from (Dosenbach et al., 2010). The maturity index prediction for the beta band (MI-beta) was obtained combining local efficiency, small world property, and resilience. The maturity index for the gamma band (MI-gamma) was obtained combining global efficiency, small world property, and resilience. The combined maturity (MI-combined) was estimated combining MI-beta and MI-gamma metrics together.

The random forest (Breiman, 2001) regression method was used for cognitive age prediction using the graph metrics found for each adjacency matrix A. One third of all available predictors were randomly sampled at each split and a total of 1000 decision trees were built. The model performance on different feature sets was estimated by Akaike information criterion and R^2 coefficient of determination, using nested bootstrapping with 1024 realizations. The reported results are as follows: for MI-beta features $R^2_{MGH} = 0.39, R^2_{OMEGA} = 0.34$ for the MGH and OMEGA data sets respectively, for MI-gamma $R^2_{MGH} = 0.48, R^2_{OMEGA} = 0.41$, and for the case of MI-combined $R^2_{MGH} = 0.52, R^2_{OMEGA} = 0.41$.

2.3. Research objectives

The analysis of the studies on the brain age prediction suggests that such modalities as MEG, structural and functional MRI convey important information related to aging. In our work we want to investigate whether the combination of these sources yields a better age prediction model for healthy subjects. In detail, we want to find how the age prediction accuracy and spread of prediction errors will change due to the incorporation of MEG data. Also, we want to examine whether the age prediction accuracy depends on the chronological age of participants. To solve this task we need to select an appropriate data set, and to construct a separate data preprocessing pipeline for every modality. Moreover, we have to find suitable machine learning techniques able to combine input data from different sources.

3. Material and methods

3.1. Cambridge Center for Aging and Neuroscience study

The data used for the brain age prediction were initially acquired as part of the Cambridge Center for Aging and Neuroscience (Cam-CAN) study, which is a multidisciplinary examination of healthy cognitive aging (Shafto et al., 2014). According to the paper, their project was focused on normal age-related changes with the aim to understand how these changes to neural structure and function interact to support cognitive abilities across the lifespan. It was implemented in 3 stages. In detail, during stage 1 a selected group of 3000 people,

aged 18 and over, were interviewed about their health and lifestyle, they underwent a core cognitive assessment, a self-completed questionnaire of lifetime experiences and physical activity. For the second stage, 700 individuals (50 men, 50 women from each age decade) were selected. Unfortunately, for the youngest decade (18–27) the data of only 56 participants (27 men, 29 women) were recorded. Different metrics of participants' cognitive health were taken, such as cognitive testing and measure of brain structure and function (MRI, fMRI, MEG), blood pressure measure, cognitive task measuring attention, language, motor and learning, memory, emotion. Finally, for the third stage 280 adults were further selected for the in-depth cognitive neuroscience assessment (attention, language, motor and learning, memory, emotion) with fMRI and MEG.

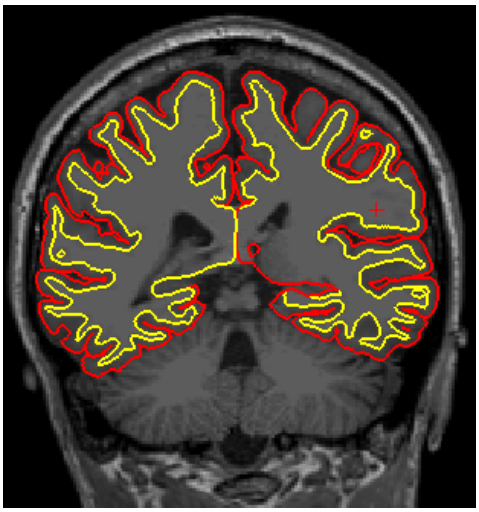
For our study we decided to focus on the information obtained during stage 2 of the Cam-CAN study (Shafto et al., 2014), since it contains raw and preprocessed structural MRI, funtional MRI (active tasks and resting state), and MEG data (active tasks and resting state) of 656 people, of cross-sectional adult lifespan (18-87 years old). We were investigating only the T1-weighted structural MRI, resting state functional MRI, and resting state MEG data. According to Taylor et al. (2017), all MRI data were acquired at the same site using a 3T Siemens TIM Trio scanner with a 32-channel head coil. Individuals rested with their eyes closed for 8 min 40 s during the acquisition of the resting state scans. All scans were saved in standard NIfTI-1.1 format using single file storage (.nii).

As it was mentioned by Taylor et al. (2017), the MEG scans were also collected at a single site using a 306 VectorView system (Elekta Neuromag, Helsinki). The scanner 102 magnetometers and 204 orthogonal planar gradiometers, it was put in a light magnetically shielded room. The sampling frequency of data was around 1 kHz with a high-pass filter of 0.03 Hz. In order to perform further artifact correction four Head-Position Indicator (HPI) coils were used to track head motion, vertical and horizontal electrooculogram (VEOG, HEOG) signals were monitoring eye blinks and movements, the electrocardiogram (ECG) signal was recording participant's pulse. During the resting state data acquisition participants were sitting with their eyes closed for at least 8 min 40 s. Neuromag's FIF format was used to store raw and maxfiltered MEG data.

3.2. MRI data preprocessing and feature engineering

The Cam-CAN Stage 2 repository contains raw and preprocessed structural and functional MRI data (Taylor et al., 2017). In our work we used the raw structural (anatomical) MRI and fMRI scans provided by the Cambridge Center for Aging and Neuroscience. Feature engineering involved different operations described in detail in the sections below.

3.2.1. Structural MRI



(a) The white (red) and pial (yellow) surfaces

(b) Subcortical volumes

Figure 2: FreeSurfer reconstruction output for subject `CC410354`. In Fig. 2a the white and pial surface are presented. The boundary between white and grey matter is shown in yellow, while the pial surface is red. Segmentation of subcortical volumes is presented in Fig. 2b.

Prior to extracting any useful information from the provided raw data, we prepared it using our custom preprocessing pipeline. In the case of anatomical images, preprocessing generally includes several operations, such as bias field correction, skull-stripping, tissue segmentation.

Bias field manifest itself as a very low-frequency variation in intensity across the MRI volume, brighter at the center and darker toward the edges of the brain. This is caused by inhomogeneities in the excitation of the head for fields with high magnetic induction $B_0 \geq 3$ T. Bias field causes problems in the analysis of structural information of MRI images. A simple approach to correct image intensity variation is to remove low-frequency signals from MRI data by a high-pass filter (Cohen et al., 2000).

Removal of the skull and other non-brain tissue in the MRI volume is called skull-stripping or brain extraction. It can be performed manually or using different automated methods (Boesen et al., 2004). The extracted brain can be obtained as part of a more general tissue segmentation step, which divides brain tissue into separate compartments (gray matter, white matter, and cerebrospinal fluid). Automatic brain tissue segmentation is a challenging task, as MRI images suffer from a variety of issues, such as noise, intensity inhomogeneity, and partial volume effect (i.e. voxels can contain a mixture of different tissue types in varying proportion). Thus, the intensity differences in T1-weighted MRI images between these tissues do not provide enough information to perform correct tissue segmentation. There are many methods proposed to tackle this problem. For example, a tissue segmentation algorithm, proposed in the work of Ashburner and Friston (2005), combines information form a probabilistic atlas with the intensity data from an image to determine the tissue class of every voxel. The atlas contains prior probability values that any voxel contains gray matter, white matter, or CSF. It can be obtained from a set of manually segmented MRI volumes.

In our work, the preprocessing of structural MRI data was implemented using the FreeSurfer software package (Fischl, 2012). It is ideal for use on large data sets, due to the fact that most of its pipelines are automated. We used the `recon-all` script that executes the FreeSurfer cortical reconstruction process for a selected subject. The reconstruction process consists of several streams: the cortical surface stream, subcortical (volume-based) stream, et al. (Dale et al., 1999; Fischl et al., 1999).

In the cortical surface stream FreeSurfer performs registration to the MNI atlas (Collins et al., 1994), bias field correction, skull stripping. Following that, it creates a model of the boundary between white matter and cortical gray matter (the white surface) as well as a model of the boundary between gray matter and cerebrospinal fluid (the pial surface). Given the models of those surfaces, the program can measure cortical surface area, cortical thickness, which is the distance between the white surface and the pial surface, and others. The volume-based stream registers MRI volumes to the MNI space, performs skull striping and assigns labels according to subcortical tissue classes using the atlas of labels provided with FreeSurfer (Fischl et al., 2002, 2004). An example of the reconstruction procedure output generated by Freesurfer for subject `CC410354` is shown in Fig. 2.

To generate features from structural MRI we processed T1-weighted volumes for every subject. Using the `mris_preproc` script provided by FreeSurfer we extracted estimates of cortical thickness and surface area resampled into the `fsaverage4` space. The data for left and right hemispheres were combined into a vector with $n_{features}(\text{thickness}) = n_{features}(area) = 5124$ entries, separately for every subject. The `asegstats2table` script was used to obtain measurements of subcortical volumes and global volume into a vector for each subject with $n_{features}(volumes) = 66$.

3.2.2. Functional MRI

Prior to any analysis fMRI data should be preprocessed. There are many approaches to solve this task, but there is a standard set of methods to choose from (Poldrack et al., 2011). The standard fMRI preprocessing stream consists of distortion correction, motion correction, slice timing correction, and spatial smoothing.

The first typical operation is distortion correction, this operation is used to combat several types of artifacts specific to gradient-echo echoplanar imaging (EPI). EPI imaging is the dominant method in fMRI studies. The inhomogeneity of the steady magnetic field B_0 due to the air-tissue interfaces causes the reduction of signal in the brain areas adjacent to these interfaces (i.e. dropout) and spatial distortions of the signal location. These artifacts typically occur near the anterior prefrontal cortex, orbitofrontal cortex, and lateral temporal lobe. Since there is no possibility to restore data from a region affected by dropout, it is suggested to reduce dropout using special MRI acquisition configurations. To correct spatial distortions one can use a field map characterizing the B_0 field (Jezzard and Balaban, 1995). A field map can be obtained using the difference in phase between the two images acquired at different echo times. This information allows to restore initial coordinates of each voxel.

During fMRI acquisition any subject will move its head, for example because of swallowing. The subject's movement causes bulk motion or a spin history effect. The first one is a movement of the head as whole. Bulk motion can significantly affect activation maps. It can be corrected using standard motion correction techniques by registering the images in the time series to a reference image. The spin history effect disrupts the fMRI signal itself. Due to the head movements, the protons change their location, which causes incorrect signal reconstruction. These can lead to large changes in the intensity of neighboring slices. This form of motion can be corrected using ICA (Jungl et al., 1998).

In most cases, fMRI data are obtained using two-dimensional MRI acquisition, in which the image is acquired one slice at a time. Owing to this, data in different slices in the volume are acquired at different times. However, during the analysis of fMRI data one assumes that all data in the image were obtained at the same time. To tackle the difference in acquisition time of different voxels, one can use slice timing correction (Henson et al., 1999). This algorithms interpolates the data in all the slices to match the timing of a selected reference slice.

Spatial smoothing of MRI volumes increases SNR ratio and reduces the variability across different subjects. It applies a filter to the image, which removes high-frequency information. A three-dimensional Gaussian filter is commonly used. The amount of smoothing imposed by a Gaussian filter is determined by the width of the distribution, which is defined as the full width of the Gaussian distribution at half-maximum (FWHM) FWHM = $2\sigma\sqrt{2\ln(2)}$, where σ is the standard deviation.

To prepare fMRI data from the Cam-CAN study we were following the functional MRI preprocessing pipeline implemented in Python, the pypreprocess package. The pipeline relies on the SPM12 software for the analysis of brain images and python-matlab interface provided by Nipype (Gorgolewski et al., 2011). The available fMRI data were visually inspected. The volumes were excluded from the study provided they had severe imaging artifacts or head movements with amplitude larger than 2 mm. After the rejection of corrupted data we obtained a subset of 626 subjects for further investigation. The fMRI volumes underwent slice timing correction and motion correction to the mean volume. Following that, co-registration between anatomical and function volumes was done for every subject. Finally, brain tissue segmentation was done for every volume and the output data were morphed to the MNI space.

To obtain functional connectomes from fMRI data the Nilearn package was used (Abraham et al., 2014). We used two parcellation atlases in our study, namely the BASC atlas with 197 functional parcels (Bellec et al., 2010) and the MODL atlas with 256 functional parcels (Mensch et al., 2016). Mean timeseries were obtained from the regions defined by each atlas. In the following step, functional connectivity matrices were estimated using either correlation between functional MRI signals or the tangent space projection (Varoquaux et al., 2010). The connectivity matrices obtained using signal correlation further underwent Fisher's r-to-z transformation. Following that, feature vectors of size either $n_{features}$(BASC 197) = 19306 or $n_{features}$(MODL 197) = 32640 were created from the lower triangle of each matrix.

3.3. Magnetoencephalography

In comparison to fMRI, magnetoencephalography provides higher temporal resolution. The Cam-CAN data set contains raw and processed MEG data. In our study we used the raw data provided in Neuromag's FIF file format. They contain information from different sensors: magnetometer, gradiometer, EEG, EOG, ECG and stimulus. Processing of MEG data is challenging because of the multi-dimensional nature of the data, and low signal-to-noise ratio (SNR). The MEG preprocessing pipeline, proposed by Jas et al. (2018), was used as a guidance during preprocessing of MEG data.

3.3.1. Cleaning channel information

We started processing of the data by cleaning channel information. Each study was visually inspected to confirm that all channels were given correct names. Due to the compatibility requirements, the coil type in the meta information of each file was initially set as 3022 or 3023. During our investigation we changed the coil type to 3024, as it is the one actually used in during the data acquisition. A list of channels with corrupted data was obtained from Elekta Mafilter log files provided by the Cambridge Center for Aging and Neuroscience. Those channels were excluded from the following processing steps.

3.3.2. Environmental artifacts

To suppress magnetic interference signal recording is typically performed in magnetically shielded rooms, the MEG systems are also equipped with gradiometers, which are less sensitive to external signal than magnetometers. Still, one needs to use advanced signal processing techniques to clean MEG recordings. In our work we utilized the signal source separation method (SSS) proposed by Taulu and Kajola (2005). This method exploits the known physical properties of magnetic fields to separate the measured signal into two linearly independent subspaces comprising external and internal, with respect to the sensor helmet, signals. It is also called signal space separation. According to this method each MEG signal can be decomposed into a linear combination of spherical harmonic functions. SNR can be increased by keeping only the first functions of such decomposition. It is generally referred to as spatial filtering.

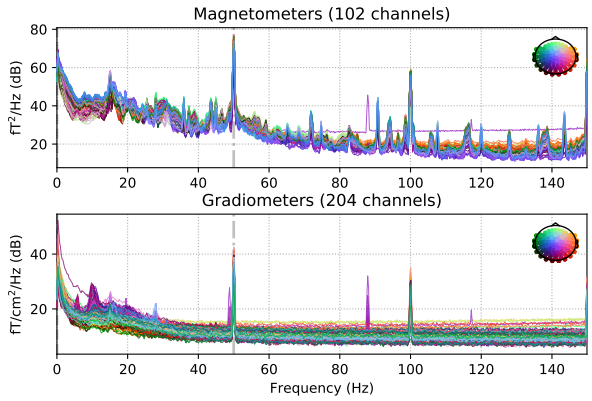
SSS requires a detailed sampling (more than about 150 channels) and a relatively high calibration accuracy, which is machine- and site-specific. In our study we used the fine-calibration coefficients and the cross-talk correction information provided by the Cambridge Center for Aging and Neuroscience.

In the real case scenario, with the data measured by an Elekta Neuromag 306-channel device, the optimal number of components is 8 for the harmonic decomposition of the internal sources, and 3 for the external sources. In this work we used the same number of spherical components and a 10 s sliding window. The correlation threshold, which is a limit between inner and outer subspacesused to reject overlapping intersecting inner/outer signals, was set to 98%. We performed no movement compensation, since there were no continuous head monitoring data available at the time of our study. The origin of internal and external multipolar moment space is fitted via head-digitization, hence specified in the head coordinate frame and the median head position during the 10 s window is used.

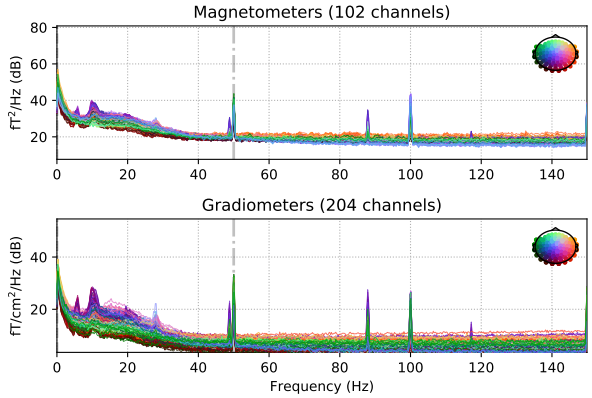
The data were processed using the MNE `maxwell_filter` function. It is important to highlight that after SSS, the magnetometer and gradiometer data

are projected from a common lower dimensional SSS coordinate system that typically spans between 64 and 80 dimensions. As a result, both sensor types contain highly similar information, which also modifies the inter-channel correlation structure (Garcés et al., 2017). Thus MNE will treat them as a single sensor type in many of the analyses that follow. A scale factor of 100 is used to bring the magnetometers to approximately the same order of magnitude as the gradiometers, as they have different units (T vs T/m).

3.3.3. Power spectral density



(a) PSD of raw data



(b) PSD of filtered data

Figure 3: Power spectral density of signals measured by magnetometers and gradiometers before (Fig. 3a) and after (Fig. 3b) application of the signal source separation method. The data was acquired for subject `CC110033` from the Cam-CAN data set. 4096 time points were used to calculate fast Fourier transform, which corresponds to a window of 12 s, given the 1 kHz sampling rate.

The spectral artifacts and bad channels can be easily detected using the power spectral density (PSD) estimates for all channels. The MNE package estimates the PSD of MEG data using Welchs method (Percival and Walden, 1993; Welch, 1967). In this method the signal for each channel is analyzed over consecutive time segments of fixed duration. The power of the discrete Fourier transform (DFT) coefficients is computed and averaged over all segments. The averaging procedure returns an unbiased and less noisy estimate of the PSD, provided that each of these segments is a realization of a stationary process. For example, in Fig. 3 PSD for the MEG data of the selected subject is shown. From the plot, it is easy to see that the SSS method removes the outliers (bad channels with abnormally high spectral power). In this study, thanks to the spatial filtering, we did not found any potentially bad channels. By inspecting a PSD plot visually one can easily find spectral artifacts. In our case, we effortlessly see the peaks from a power line at 50 Hz. This particular artifact should not affect the age prediction since it is the same for every subject, thus we did not remove it from the data.

3.3.4. Temporal filtering

To further improve SNR of MEG data one can exclude signals with frequencies that are outside of the band where the responses energy is supposed to lie. To distinguish it from spatial filtering, it is referred to as time-domain or temporal filtering. In our work, the resting state brain responses were band-pass filtered the range of frequencies from 0.1 Hz up to 150 Hz.

3.3.5. Filtering of bad data segments

In some case the filtered data still may include bad data segments and bad epochs due to transient jumps. The MNE package provides means to remove bad trials by analyzing their peak-to-peak amplitude. Specifically, it discards any trial which peak-to-peak amplitude exceeds a certain subject-specific rejection threshold. Additionally, instead of selecting the threshold manually for each subject, we used the autoreject algorithm (Jas et al., 2017). Autoreject is an unsupervised algorithm, which find the optimal threshold by minimizing the Frobenius norm between the average signal of the training set and the median signal of the validation set.

3.3.6. Physiological artifacts

Physiological artifacts in MEG data are the ones caused by subject's heart beats and eye blinks. MEG recording systems typically provide this physiological information in the form of EOG and ECG signal recordings. In our work we mitigated the influence of physiological artifacts using the signal space projection (SSP) method (Uusitalo and Ilmoniemi, 1997). This method uses the concept of signal and interference subspaces, similar to the source space separation algorithm. According to the SSP method, the measurement data can be projected onto a hyperplane orthogonal to the inference subspace, thus completely removing the contribution of the unwanted subspace.

In detail, we excluded bad data segments from the EOG/ECG channels using global autoreject (Jas et al., 2017). Then EOG/ECG signals were used to detect time points associated with artifacts in the MEG signals. To isolate segments dominated by artifacts we extracted epochs using 500 ms windows around those

events and averaged those epochs. For each subject data we obtained a vector $\mathbf{x} \in \mathbb{R}^{n_s \times t_e}$, where n_s is the number of sensors and t_e the number of time points. For each of those vectors the artifact amplitude dominated signal amplitude. Assuming that the signal space and the artifact space were orthogonal to each other, we extracted the artifact from real signal by computing the eigenvalue decomposition of the covariance matrix $\mathbf{x}\mathbf{x}^\top \in \mathbb{R}^{n_s \times n_s}$ and removing the first eigenvector (SSP vector) for each sensor type and each artifact type. This procedure dampens the artifact influence, while not removing it completely.

3.3.7. Epoching

In event-related paradigm, a stimulus channel contains binary-coded trigger pulses to mark the onset/offset of events. There are no proper events of interest in the resting state paradigm, so we used fixed length segments of 30 s of data with no overlap. We extracted segments of data from the continuous recording around events of interest and stored them as single trials, which are also called epochs in MNE. The duration of time windows was selected to have a higher frequency resolution. The selected overlap time allowed to get a less noisy estimate of the brain dynamics in resting state.

3.3.8. Supervised spatial filtering

Following prepossessing of MEG signals, one needs to generate suitable input data for machine learning models. To find the relationship between aging and MEG data Khan et al. (2018) analyzed functional connectivity using graph theory. In our work we decided to follow a simpler path, and to rely solely on signal processing techniques. Thus, we inspected MEG signals for age-related patterns using the source power comodulation (SPoC) algorithm, introduced by Dähne et al. (2014). It is a supervised spatial filtering algorithm that uses a target variable to guide the signal decomposition process. Incorporating this information allows the algorithm to give preference to components whose power comodulates with the target variable. SPoC chooses spatial filters to maximize the covariance between the power of the filtered signals and the target variable, which is the chronological age in our case. In our work, a covariance matrix of the MEG signals measured by magnetometers was calculated separately for every subject in these frequency bands: 0.1–1.5 Hz (low), 1.5–4.0 Hz (delta), 4.0–8.0 Hz (theta), 8.0–15.0 Hz (alpha), 15.0–26.0 Hz (low beta), 26.0–35.0 Hz (high beta), 35.0–50.0 Hz (low gamma), 50.0–74.0 Hz (medium gamma), and 76.0–120.0 Hz (high gamma). These matrices were processed using SPoC, while subject's age was set as the target variable. We utilized the spatial filters learned from the training data to estimate features for age prediction.

3.4. Machine learning

3.4.1. Cognitive age regression

It is easy to see that the task of cognitive age prediction given processed MRI, MEG data is a typical supervised learning problem (Hastie et al., 2009). The provided data can be denoted as a vector $x = (x_1, x_2, \ldots, x_p)^\top$, where p is the number of predictors or features. In our problem we want to predict values of the subjects' age $y \in \mathbb{R}_+$, i.e. a target variable, which should belong to a set of continuous, real, non-negative numbers. In this case regression methods should be utilized.

Linear models are widely used to solve the task of decoding, i.e. prediction of behavior or biomarkers from brain images or signals (Varoquaux et al., 2017). This is due to the fact that stable but biased linear models are not only easy to train or interpret, but they are less affected by the curse of dimensionality (Bellman, 1961), than more flexible prediction models given the dimensions of neuroimaging problems. Thus we decided to use a linear regression model, which supposes that the target variable y can be predicted from a linear combination of the features (Hastie et al., 2009). It can be written as

$$y = \beta_0 + \sum_{j=1}^{p} x_j \beta_j = \beta_0 + x^\top \beta \tag{1}$$

where $\beta \in \mathbb{R}$ is a vector of coefficients, $\beta_0 \in \mathbb{R}$ is the intercept or bias. The latter term can be included in the vector of coefficients, thus Eq. 1 becomes

$$y = x^\top \beta \tag{2}$$

where $\beta \in \mathbb{R}^{(p+1)}$ and the constant variable 1 was included in x.

Assume we were given a learning set or training data as in Eq. 3, where x_i is an input vector and y_i is the corresponding response.

$$\mathcal{D}_n = \{(y_i, X_i),\ i = 1, \ldots, n\} \tag{3}$$

There are many methods to fit the linear model defined in Eq. 2 to the given data, the most popular among them is the method of least squares (Hastie et al., 2009). According to this method, the coefficients β are selected so that the residual sum of squares RSS will be minimal

$$\mathrm{RSS}(\beta) = (y - X\beta)^\top (y - X\beta) \tag{4}$$

where X is an $n \times p$ matrix where each row is an input vector, and y is an n-vector of the outputs in the training set. Differentiating Eq. 4 with respect to β we obtain

$$\frac{\partial \mathrm{RSS}}{\partial \beta} = -2X^\top (y - X\beta) \tag{5}$$

$$\frac{\partial \mathrm{RSS}}{\partial \beta \partial \beta^\top} = 2X^\top X \tag{6}$$

If X has full column rank, i.e. input vectors are linearly independent, then $X^\top X$ is positive definite. Thus we can find a minimum value for RSS by finding the point where the first derivative is zero

$$X^\top(y - X\beta) = 0 \tag{7}$$

The equation above has the unique solution

$$\hat{\beta} = \left(X^\top X\right)^{-1} X^\top y \tag{8}$$

where $\hat{\beta}$ is a vector of estimated coefficients of the linear model defined in Eq. 2.

Typically, the least squares estimates $\hat{\beta}$ have low bias but large variance, and use a large number of predictors. Prediction accuracy and the robustness to overfitting might be improved by introducing some bias into the model (Hastie et al., 2009). This can be achieved using variable subset selection, shrinkage methods (ridge regression, the lasso), partial least squares, and principal components regression. In the work of Frank and Friedman (1993) a detailed study of those methods can be found. The authors state that ridge regression (RR) gives the minimal prediction error in comparison to the other methods. Moreover, RR shrinks coefficients smoothly, rather than in discrete steps (Hastie et al., 2009). Thus we decided to use this method in our work.

Ridge regression was initially introduces in the work of Hoerl and Kennard (1970). It imposes a penalty on the size of coefficients. The ridge coefficient estimates $\hat{\beta}_{ridge}$ minimize a penalized residual sum of squares

$$\mathrm{RSS}(\lambda) = (y - X\beta)^\top(y - X\beta) + \lambda\beta^\top\beta \tag{9}$$

Differentiating this equation with respect to β we obtain

$$\frac{\partial \mathrm{RSS}}{\partial \beta} = -2X^\top(y - X\beta) + 2\lambda\beta \tag{10}$$

$$\frac{\partial \mathrm{RSS}}{\partial \beta \partial \beta^\top} = 2X^\top X + 2\lambda \tag{11}$$

By utilizing the same approach as for finding the least squares estimates in Eq. 8, we can find the final solution for the ridge regression coefficient estimates Eq. 12, where I is the $p \times p$ identity matrix and $\lambda \geq 0$ is the regularization parameter that controls the trade-off between data-fitting and regularization. It is easy to see that such regularization adds a positive constant to the diagonal of $X^\top X$, thus it makes the problem nonsingular even for rank deficient $X^\top X$. There are multiple strategies to select the value of λ. In our work we decided to use cross-validation, we were looking for the best value of λ across $n = 100$ points on a logarithmic scale between $a = 10^{-3}$ and $b = 10^5$, which is a set of points $S = \{x_1 = a,\ x_{i+1} = x_i r^{(b-a)/n}\}_{i=1}^{n}$, where $r = 10$.

$$\hat{\beta}_{ridge} = (X^\top X + \lambda I)^{-1} X^\top y \tag{12}$$

3.4.2. Stacking regressions

According to Breiman (1996), stacking regressions is a method for forming linear combinations of different predictors to give improved prediction accuracy. To calculate the coefficients in the combination it uses cross-validation and least squares under non-negativity constraints. The stacking procedure was initially proposed by (Wolpert D., 1992), it implies that given a set of prediction functions $f_1(x), \ldots, f_k(x)$ a more accurate predictor function can be built by combining those. The method for combination is based on level 1 data defined as follows: leave out the ith case and repeat the procedures for constructing the predictor functions, getting $f_j^{(-i)}(x),\ j = 1, \ldots, k$. We define a vector $z_i \in \mathbb{R}^k$ by

$$z_{ij} = f_j^{(-i)}(x_i) \tag{13}$$

Then the data on level 1 is $(y_i, z_i),\ i = 1, \ldots, n$. This data is typically used for selecting one of the f_j with respect to some minimization criterion, e.g. least squares. According to Wolpert D. (1992), the level 1 data contains more information and can be used to construct a combination of the f_j that exhibits even better accuracy. A simple approach to solve the problem of combining the f_j is to consider a linear combination as in Eq. 14 (Breiman, 1996)

$$f(x) = \sum_j \alpha_j f_j(x) \tag{14}$$

where $\alpha \in \mathbb{R}^k$ such that

$$\alpha = \arg\min_\alpha \sum_i (y_i - \alpha z_i)^2 \tag{15}$$

Here the level 1 data is generated using leave-one-out cross-validation (LOOCV), although the suitable level 1 data can be successfully obtained by less computationally demanding k-fold cross validation (Breiman, 1996).

To perform stacking one can use not just linear regression as in Eq. 15. For example in the work of Rahim et al. (2016), the stacking step was implemented using either logistic regression, or ridge regression, or random forests. They found that three classifiers led to the similar improvements in prediction accuracy, though random forests gave more stable output than logistic regression and ridge classifier, i.e. the standard deviation of prediction accuracy was significantly reduced. According to the results of (Liem et al., 2017) stacking with random forests significantly improved prediction accuracy relatively to the originally used classifiers.

We combined single-source using RF, a flow-chart depicting our stacking approach can be seen in Fig. 4.

3.4.3. Random forests

The bootstrap is a general tool for assessing statistical accuracy (Hastie et al., 2009). Assuming we have a model fit to a set of training data as in Eq. 3, we

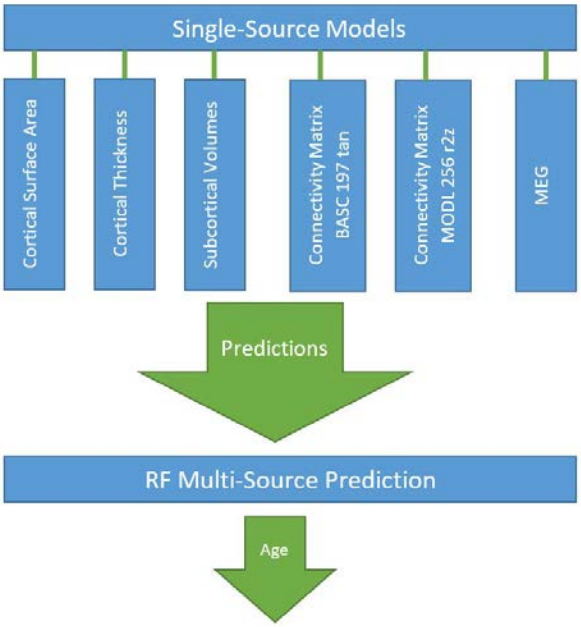


Figure 4: Stacking diagram. The predictions from selected single-source models are stacked using the random forest algorithm to obtain age values.

should randomly draw data sets with replacement from the training data m times, each sample the same size as the original training set. In this way we will produce m bootstrap data sets $\mathcal{D}_n^{*b}$, $b = 1, 2, \ldots, m$. In the following step, we should fit the model again to each of the bootstrap data sets, and examine the behavior of the fits over those sets. Bootstrap aggregation or bagging finds the average prediction of the model over the collection of bootstrap samples $\mathcal{D}_n^{*b}$ from our training data $\mathcal{D}_n$, see Eq. 16.

$$\hat{f}_{bag} = \frac{1}{m} \sum_{b=1}^{m} \hat{f}^{*b}(x) \tag{16}$$

Here $\hat{f}^{*b}(x)$ is the fit of the model to $\mathcal{D}_n^{*b}$. Bagging averages the outputs of noisy unbiased models, thus it reduces the variances relative to any of those models when considered separately.

Regression trees partition the space of all predictor variable values into k disjoint regions R_j, $j = 1, 2, \ldots, k$ using greedy, top-down recursive binary splitting. The jth terminal node of the tree corresponds to a region R_j. A constant γ_j is assigned to each such region, which is typically the mean or the mode of the training observations in the corresponding region. This constant is called terminal-node value (Hastie et al., 2009). For every observation x that falls into the region R_j we make the same prediction by returning γ_j. Hence, we can write it as

$$T(x; \Theta) = \sum_{j=1}^{k} \gamma_j I(x \in R_j) \tag{17}$$

where $I(x \in R_j) = 1$ if $x \in R_j$ and $I(x \in R_j) = 0$ otherwise, $\Theta = \{R_j, \gamma_j\}_1^m$, m is selected empirically.

The random forests algorithm developed by Breiman (2001) combines decision trees and bagging. The idea behind this method is to improve the variance reduction of bagging by decreasing the correlation between the trees. It builds a large ensemble of de-correlated decision trees $\{T_b\}_1^m$ over bootstrap samples $\mathcal{D}_n^{*b}$. To reduce the correlation between those trees one should randomly select the input variables during the tree growing process. In detail, when growing a tree T_b on a bootstrapped data set $\mathcal{D}_n^{*b}$, at each split only a randomly selected subset of size $s <= p$ of the input variables is considered. In most cases, $s = \sqrt{p}$. The final random forest regression predictor is shown in Eq. 18.

$$\hat{f}_{rf}^{m}(x) = \frac{1}{m} \sum_{b=1}^{m} T(x, \Theta_b) \tag{18}$$

In our work, the values of hyperparameters were the same as the ones used by Liem et al. (2017). In detail, we were building random forest using $m = 100$ estimators, while all the available features were considered when looking for the best split $s = p$, the default values of the other hyperparameters were used, see the `RandomForestRegressor` class from scikit-learn (Pedregosa et al., 2011). Models were trained and evaluated using cross-validation.

3.4.4. Model evaluation

While solving statistical learning problems we want to find the most suitable model among the available ones, which is the model selection problem. At the same time, we want to assess the performance of our estimated model $\hat{f}$ (its generalization error) on independent test data, i.e. the model assessment problem. The generalization or test error can be approximated using training error, see Eq. 19, which is equals the average loss over the training data $\mathcal{D}_n$. The loss can be quantified using the loss function L, which is typically either a squared or absolute value of a residual, see Eq. 20.

$$\overline{\text{err}} = \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{f}(x_i)) \tag{19}$$

$$L(y_i, \hat{f}(x_i)) = \begin{cases} (y_i - \hat{f}(x_i))^2 \\ |y_i - \hat{f}(x_i)| \end{cases} \tag{20}$$

Unfortunately, this is not a good estimate, due to the fact that a model with high complexity may overfit the data. As a result, it will have very low training error, but high test error.

One can solve those problems by randomly dividing the data set into three parts: a training set, a validation set, and a test set. Typically, the given data are shuffled, then 50% of the samples are allocated for the training set, used to fit the models to. The other 50% are equally split into the validation set, used to find prediction error

for model selection, and the test set, used to estimate the generalization error of the selected model. The splits should preserve the ratio of samples between groups, i.e. stratified splitting. This approach is suitable when the amount of data is relatively large. For other cases, it is recommended to use K-fold or leave-one-out cross-validation (Hastie et al., 2009).

K-fold cross-validation randomly splits the available data into k roughly equal-sized parts. jth part of those splits is used as the validation set, while other $k-1$ parts are preserved to fit the model. The process is repeated k times and the performance scores from all runs are averaged. If we use a function $\kappa : \{1, \ldots, n\} \mapsto \{1, \ldots, k\}$ indicating the partition to which observation i is allocated by the randomization, then the prediction error can be found as

$$CV(\hat{f}) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{f}^{-\kappa(i)}(x_i)) \tag{21}$$

It is recommended to select $k = 5$ or $k = 10$. In case the amount of available data is particularly scarce one can choose $k = n$, which is called leave-one-out cross-validation (LOOCV).

Besides test error, once can estimate how well the model fits given data using the R^2 statistic (the coefficient of determination), which equals to the proportion of explained variance (James et al., 2013)

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} \tag{22}$$

where RSS is defined as in Eq. 4 and $\text{TSS} = \sum_{i=1}^{n}(y_i - \overline{y})^2$ is the total sum of squares, and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the mean of the observed data. The R^2 metric measures the proportion of variability in y that can be explained by the model using observations x.

3.5. Implementation details

During the inspection of the raw MRI, fMRI, and MEG data provided by the Cambridge Center for Aging and Neuroscience, we selected 625 subjects with suitable MRI, fMRI studies and 640 subjects with satisfactory MEG records. The regression models were trained on 588 studies from the CamCAN data set, which contained high quality information for every modality.

The MEG, structural and functional MRI features were extracted for every subject by Python scripts. Single-source models were trained on the structural MRI features (i.e. cortical surface area, cortical thickness, and subcortical volumes), the functional MRI features, and MEG data. The fMRI features were obtained from connectivity matrices estimated either in the tangent space or using Fisher's r-to-z transformation (corresponding shortcuts `tan` and `r2z`), on brain parcellations using the BASC (197 functional parcels) or MODL (256 functional parcels) atlas. Multi-source models were trained on different combinations of features, such as structural MRI features; structural and functional MRI features; MEG and structural MRI features; MEG and functional MRI features; MEG, fMRI and structural MRI.

To build regression models of the subject's age the `RidgeCV`, `RandomForestRegressor` classes from `scikit-learn` (Pedregosa et al., 2011) were utilized. Stacking was added using the code from a `scikit-learn` feature proposal. Learning curves were obtained using the `learning_curve` function. 10-fold cross-validation was selected to evaluate every model and find its age predictions. The `cross_val_predict` function allowed us to find the predicted age for every participant. The performance of every model was estimated via mean absolute error and the coefficient of determination using the `cross_val_score` function.

4. Results

The objective of the current investigation was to develop a method for the brain age prediction that combines information from MEG, structural MRI and functional connectivity data. To achieve this aim we stacked regularized linear models (ridge regression) with the random forest algorithm. To evaluate the performance of the age prediction models we estimated mean absolute prediction error and the coefficient of determination. MAE and its distribution can be seen in Fig. 5, corresponding values and R^2 scores are presented in Table A.1. All age regression models exhibit good accuracy, with MAE ranging from 4.75 to 10.39 years and R2 ranging from 0.48 to 0.89. Multi-source models perform better than single-source models, stacking with MEG data improves the age prediction accuracy. The highest accuracy was achieved by the model combining MEG, structural and functional MRI. The second best model was the one combining structural and functional MRI. Although stacking of single-source models improves prediction accuracy and the coefficient of determination in comparison to the used single-source models, the standard deviation does not follow this pattern and shows less straightforward behavior.

For every model we also obtained a scatter plot of the chronological age versus the age predicted by the model for every participant, and its learning curves. For example, in Fig. 6 these plots were done for the multi-source model stacking MEG, structural MRI, and functional MRI data. In Fig. 6a we see no outliers, which means that our preprocessing pipeline was run successfully. We performed this visual inspection for every model. Learning curves in Fig. 6b show that MAE decreases with the addition of new training samples, the same behavior was observed for other models, which implies that prediction accuracy can be further improved by adding new studies.

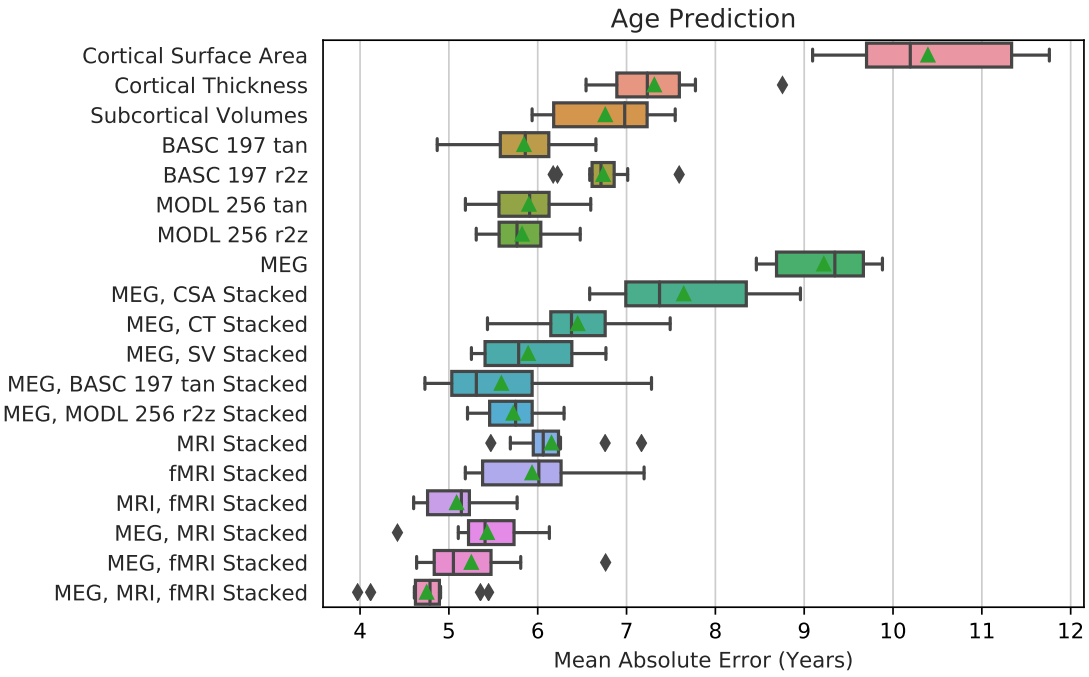


Figure 5: MAE for single-source models and the models obtained via stacking of single source models. `CSA`, `CT`, and `SV` are abbreviations of cortical surface area, cortical thickness, and subcortical volumes correspondingly. FMRI features estimated in the tangent space are denoted using the `tan` abbreviation, while `r2z` denotes fMRI features calculated using Fisher's r-to-z transformation. `MRI Stacked` combines all structural features, `fMRI Stacked` combines the MODL 256 r2z and BASC 197 tan fMRI features. A green triangle and a vertical black line in every interquartile range correspond to the mean and median of the distribution. Multi-source models (`Stacked`) perform better than single-source models, stacking with MEG data improves the age prediction accuracy. The highest accuracy was achieved by the model combining MEG, structural and functional MRI, namely the `MEG, MRI, fMRI Stacked` model.

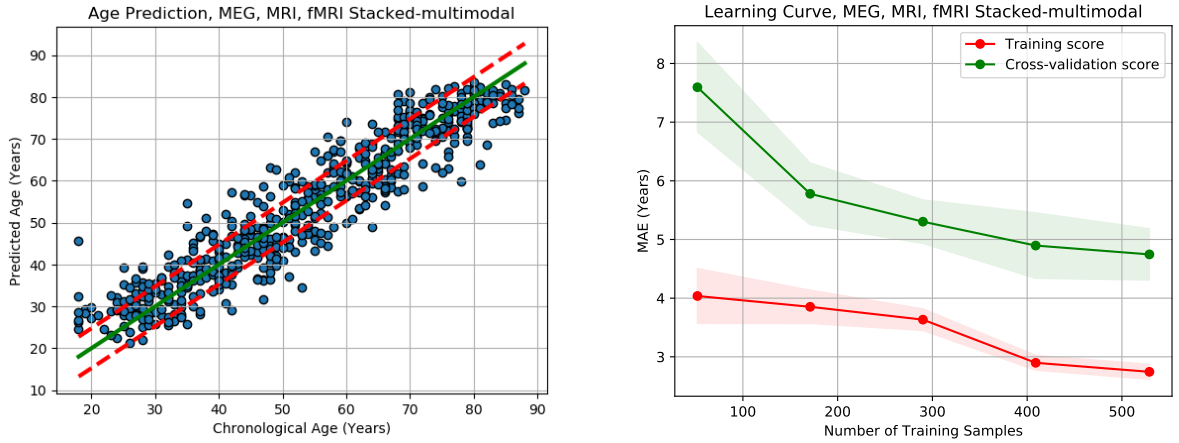
In addition, we estimated mean absolute error of age prediction of all models for each age group, where a group constitutes of participants with age difference not more than 10 years. In total, we obtained 7 age groups: 18–28 years, 28–38 years etc. The dependency of MAE on age groups can be seen in Fig. 7. Singles-source models trained on cortical surface area and MEG data exhibit strong U-shaped dependency dependency between mean absolute prediction error and the chronological age, which indicates that the age prediction errors were higher for younger and older participants. For the other models the dependency is weaker, and the multi-source models exhibit the least variation of MAE across age groups.

5. Discussion

In our work we investigated whether the combination of MEG, structural and functional MRI data yields a better age prediction model for healthy subjects. As we anticipated, the best age prediction accuracy was achieved by the multi-source model combining MEG, structural and functional MRI data with MAE = 4.75 years and R^2 = 0.89. The accuracy of the model trained on structural and functional MRI data was MAE = 5.09 years and R^2 = 0.87, the models trained only on anatomy or functional connectivity data reached MAE = 6.16 years and R^2 = 0.82, MAE = 5.94 years and R^2 = 0.82 respectively. Based on this results we can see that by integrating information from different modalities we improved prediction accuracy by approximately one year, in comparison to the structural or functional MRI only models. Addition of MEG data to the `MRI, fMRI Stacked` model increased prediction accuracy by approximately 0.3 years. In contrast to the prediction accuracy of the investigated models, the standard deviation of the age prediction did not show consistent improvement. The increase in prediction accuracy of stacked models is due to the addition of new information by each modality.

Among the single-source models trained on structural MRI data, the ones trained on subcortical volumes showed the best value of MAE, which agrees with the brain aging patterns discovered earlier by Good et al. (2001); Pfefferbaum et al. (1994). The performance of the models based on cortical thickness information was almost as good, while linear models exploiting cortical surface area showed significantly lower accuracy. This observation is in agreement with the study of Storsve et al. (2014), where a clear correlation between aging and cortical thinning was found, while the relationship between cortical surface area and brain maturing was less pronounced.

While doing stacking of functional MRI single-



(a) (b)

Figure 6: The age prediction model combing data of all modalities (MEG, structural MRI, and functional MRI). The chronological age and age predicted by this model for every subject are shown in Fig. 6a. Separate subjects are represented by blue circles, the perfect prediction is the solid green line, dashed red lines bound values that differ from the perfect prediction by less than one standard deviation. The investigated model captures the relationship between the chronological age and data quite good, demonstrating uniform variance and no outliers. The learning curve of the same model is presented in Fig. 6b, the inverse dependency between the number of training sample (i.e. participant's data) and MAE implies that prediction accuracy can be further improved by adding new studies.

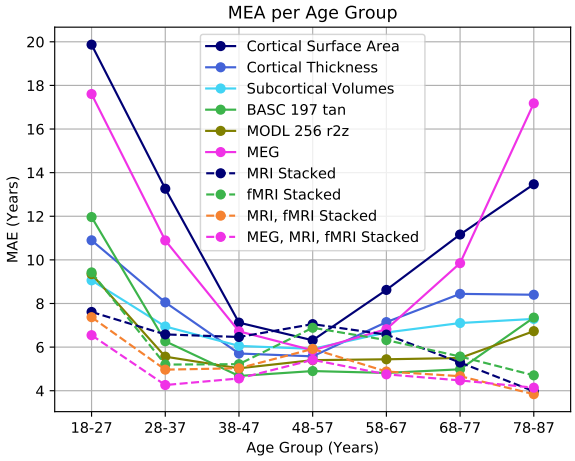


Figure 7: MAE depending on age groups. A group is a set of participants whose age difference is not more than 10 years. Straight lines correspond to single-source models, while dashed lines correspond to multi-source models obtained by stacking. Multi-source models (`Stacked`) exhibit the least variation of MAE across age groups.

source models, we selected a connectivity matrix estimated for the BASC atlas in the tangent space (`BASC 197 tan`) and another one estimated for the MODL atlas using r-to-z transformation (`MODL 256 r2z`), due to the fact that corresponding single-source models showed the best performance among the other models trained on fMRI data and were less correlated than other functional MRI-based features, see Fig. A.8. The other features (anatomy, fMRI, MEG) did not show any significant correlation between each other.

It is important to note that stacking of MEG data with other modalities substantially improves the brain age prediction accuracy. In Fig. 5 one can see how adding MEG data to any of the structural and functional MRI features affects prediction accuracy. Moreover, the best age prediction accuracy was obtained for stacked multimodal regression function trained on MEG, structural and functional MRI features. The improvement introduced by MEG data into the age prediction accuracy of this multi-source model was not very large, 0.3 years in comparison to the multi-source model trained on fMRI and structural MRI data. This is because the single-source model trained on MEG was not capable of achieving high accuracy on its own, its best score MAE = 9.22 years.

As one can see in Fig. 7, mean absolute error of the brain age prediction depends on the subject's age. High variations in MAE values depending on the chronological age were observed for single source models trained on cortical surface area and MEG. This models tend to have higher prediction accuracy for the young and old participants than for the middle-aged participants, which implies that the relationship between the participant's age and these features is non-linear. In contrast, the models trained on fMRI features tend to have significantly lower variation in MAE values for different age groups. Multi-source models demonstrated notable decrease in MAE variance depending on the chronological age, which is caused by the fact that aggregation via random forest lowers the bias in the model parameter estimation and better captures non-linear relationships in the training data, for example see Fig. 6a, 7.

There is direct correspondence between the performance of the models trained on structural and functional MRI in our study (Fig. 1) and the results presented in

the work of Liem et al. (2017). In detail, the accuracy demonstrated by the stacking of the fMRI and anatomical MRI features is MAE = 4.29 years in the work of Liem et al. (2017) and MAE = 5.09 years in our study. In general, achieved MAE is higher in comparison to the results in the paper mentioned above, this may be due to the fact that we were using a different data set, which is almost four times smaller than the LIFE data set (i.e. 656 versus 2600 unique studies). The performance of single-source models trained on MEG data was in accordance with the work of Khan et al. (2018). We achieved $R^2 = 0.58$ for the participants aged 18–87 years, while the scores in the work mentioned above are $R^2_{MGH} = 0.52, R^2_{OMEGA} = 0.41$ for the participants aged 7–29 years.

In the work of Liem et al. (2017), it was demonstrated that head motion had no significant influence on the brain age prediction. Moreover, the authors found that exclusion of motion-related signals affected meaningful variance related to age. Thus in our work we did not consider motion confounds while working with MRI and MEG data. Though, it is interesting to see whether for MRI and MEG data from the Cam-CAN data set motion correction can be of any use. The prediction accuracy of the MEG single-source model and related models can be improved by using more complex preprocessing pipeline of the MEG signals. For example, instead of considering signals in the sensor space one can analyze them in the source space. Although, prediction accuracy can be further improved by adding new studies, as can be seen from learning curves in Fig. 6b, it is hard to find a data set with large amount of MEG, structural and functional MRI data. Combining data from different sites might be of interest. However, as it was shown by Liem et al. (2017), usage of data from different sites substantially lowers the age prediction accuracy. Nevertheless, it is unclear how much the age prediction error can be lowered. Some prediction error will always persist because of individual differences in the human brains of the same age (Liem et al., 2017).

The main focus of this study was to assess how incorporation of MEG, fMRI and anatomy data can improve the brain age prediction. Though, to incorporate brain age as a reliable biomarker for neurological and psychiatric diseases we need to understand better how the models combining MEG and MRI data will behave in relation to the diseased brains. Regarding the fact that structural and functional MRI data showed promising results (Cole and Franke, 2017; Franke and Gaser, 2012; Franke et al., 2010; Liem et al., 2017), the combination of MEG signals with structural and functional MRI data should further advance the research in this area.

6. Conclusions

In the present study we determined that incorporation of MEG, functional connectivity and structural MRI data can improve the brain age prediction accuracy. To achieve this we utilized stacking of the ridge regression models using the random forest algorithm, this approach increased the prediction accuracy of multi-source models compared to single-source models. We demonstrated that the improvement was primarily due to the fact that stacking takes into account non-linear dependencies present in the training data. The further investigations are required to be able to employ the predicted brain age as a reliable biomarker.

7. Acknowledgments

I would like to express sincere gratitude to my scientific supervisors, Alexandre Gramfort and Denis Engemann, for their invaluable suggestions and support. I am very grateful to David Sabbagh, Gael Varoquaux, and for the other members of the Parietal team.

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., Varoquaux, G., 2014. Machine learning for neuroimaging with scikit-learn. Frontiers in neuroinformatics 8, 14.

Ashburner, J., Csernansk, J.G., Davatzikos, C., Fox, N.C., Frisoni, G.B., Thompson, P.M., 2003. Computer-assisted imaging to assess brain structure in healthy and diseased brains. The Lancet Neurology 2, 79–88.

Ashburner, J., Friston, K.J., 2005. Unified segmentation. NeuroImage 26, 839–851.

Atkinson A.J., J., Colburn, W.A., DeGruttola, V.G., DeMets, D.L., Downing, G.J., Hoth, D.F., Oates, J.A., Peck, C.C., Schooley, R.T., Spilker, B.A., Woodcock, J., Zeger, S., 2001. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. Clinical Pharmacology & Therapeutics 69, 89–95.

Bellec, P., Rosa-Neto, P., Lyttelton, O.C., Benali, H., Evans, A.C., 2010. Multi-level bootstrap analysis of stable clusters in resting-state fMRI. NeuroImage 51, 1126–1139.

Bellman, R.E., 1961. Adaptive Control Processes. Princeton University Press .

Boesen, K., Rehm, K., Schaper, K., Stoltzner, S., Woods, R., Lüders, E., Rottenberg, D., 2004. Quantitative comparison of four brain extraction algorithms. NeuroImage 22, 1255–1261.

Breiman, L., 1996. Stacked regressions. Machine Learning 24, 49–64.

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32.

Bzdok, D., 2017. Classical Statistics and Statistical Learning in Imaging Neuroscience. Frontiers in Neuroscience 11, 1–23.

Castellanos, F.X., Di Martino, A., Craddock, R.C., Mehta, A.D., Milham, M.P., 2013. Clinical applications of the functional connectome. NeuroImage 80, 527–540.

Cohen, M.S., DuBois, R.M., Zeineh, M.M., 2000. Rapid and effective correction of RF inhomogeneity for high field magnetic resonance imaging. Human Brain Mapping 10, 204–211.

Cole, J.H., Franke, K., 2017. Predicting Age Using Neuroimaging: Innovative Brain Ageing Biomarkers. Trends in Neurosciences 40, 681–690.

Cole, J.H., Leech, R., Sharp, D.J., 2015. Prediction of brain age suggests accelerated atrophy after traumatic brain injury. Annals of Neurology 77, 571–581.

Collins, L., Neelin, P., M. Peters, T., Evans, A., 1994. Automatic 3D Intersubject Registration of MR Volumetric Data in Standardized Talairach Space. Journal of computer assisted tomography 18, 192–205.

Dähne, S., Meinecke, F.C., Haufe, S., Höhne, J., Tangermann, M., Müller, K.R., Nikulin, V.V., 2014. SPoC: A novel framework for relating the amplitude of neuronal oscillations to behaviorally relevant parameters. NeuroImage 86, 111–122.

Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical Surface-Based Analysis. Segmentation and Surface Reconstruction. NeuroImage 9, 179–194.

Davatzikos, C., Xu, F., An, Y., Fan, Y., Resnick, S.M., 2009. Longitudinal progression of Alzheimer's-like patterns of atrophy in normal older adults: the SPARE-AD index. Brain 132, 2026–2035.

Dennis, E.L., Thompson, P.M., 2014. Functional Brain Connectivity Using fMRI in Aging and Alzheimer's Disease. Neuropsychology Review 24, 49–62.

DoEaSA-P, 2017. World Population Ageing 2017. Technical Report. United Nations. New York, NY, USA.

Dosenbach, N.U., Nardos, B., Cohen, A.L., Fair, D.A., Power, J.D., Church, J.A., Nelson, S.M., Wig, G.S., Vogel, A.C., Lessov-Schlaggar, C.N., Barnes, K.A., Dubis, J.W., Feczko, E., Coalson, R.S., Pruett, J.R., Barch, D.M., Petersen, S.E., Schlaggar, B.L., 2010. Prediction of individual brain maturity using fMRI. Science 329, 1358–1361.

Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A., Vapnik, V., 1997. Support vector regression machines. Advances in Neural Information Processing Systems 9: Proceedings of the 1996 Conference 1, 155–161.

Fischl, B., 2012. FreeSurfer. NeuroImage 62, 774–781.

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M., 2002. Whole Brain Segmentation. Neuron 33, 341–355.

Fischl, B., Sereno, M.I., Dale, A.M., 1999. Cortical Surface-Based Analysis. Inflation, Flattening, and a Surface-Based Coordinate System. NeuroImage 9, 195–207.

Fischl, B., Van Der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D.H., Busa, E., Seidman, L.J., Goldstein, J., Kennedy, D., Caviness, V., Makris, N., Rosen, B., Dale, A.M., 2004. Automatically Parcellating the Human Cerebral Cortex. Cerebral Cortex 14, 11–22.

Fisher, R.A., 1921. On the probable error of a coefficient of correlation deduced from a small sample. Metron 1, 3–32.

Frank, I.E., Friedman, J.H., 1993. A Statistical View of Some Chemometrics Regression Tools. Technometrics 35, 109–135.

Franke, K., Gaser, C., 2012. Longitudinal Changes in Individual BrainAGE in Healthy Aging, Mild Cognitive Impairment, and Alzheimer's Disease 1Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.e. GeroPsych 25, 235–245.

Franke, K., Ziegler, G., Klöppel, S., Gaser, C., 2010. Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters. NeuroImage 50, 883–892.

Garcés, P., López-Sanz, D., Maestú, F., Pereda, E., 2017. Choice of Magnetometers and Gradiometers after Signal Space Separation. Sensors 17, 2926.

Good, C.D., Johnsrude, I.S., Ashburner, J., Henson, R.N., Friston, K.J., Frackowiak, R.S., 2001. A Voxel-Based Morphometric Study of Ageing in 465 Normal Adult Human Brains. NeuroImage 14, 21–36.

Gorgolewski, K., Burns, C.D., Madison, C., Clark, D., Halchenko, Y.O., Waskom, M.L., Ghosh, S.S., 2011. Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. Frontiers in Neuroinformatics 5.

Hansen, P.C., Kringelbach, M.L., Salmelin, R. (Eds.), 2010. MEG: An Introduction to Methods. Oxford University Press, Inc.

Hastie, T., Tibshirani, R., Friedman, J.H., 2009. The elements of statistical learning: data mining, inference, and prediction, 2nd Edition. Springer series in statistics, Springer.

Henson, R., Buechel, C., Josephs, O., Friston, K., 1999. The slice-timing problem in event-related fMRI. NeuroImage 9.

Hipp, J.F., Hawellek, D.J., Corbetta, M., Siegel, M., Engel, A.K., 2012. Large-scale cortical correlation structure of spontaneous oscillatory activity. Nature Neuroscience 15, 884–890.

Hipp, J.F., Siegel, M., 2015. BOLD fMRI correlation reflects frequency-specific neuronal correlation. Current Biology 25, 1368–1374.

Hoerl, A.E., Kennard, R.W., 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 12, 55–67.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning. volume 103 of *Springer Texts in Statistics*. Springer New York, New York, NY.

Jas, M., Engemann, D.A., Bekhti, Y., Raimondo, F., Gramfort, A., 2017. Autoreject: Automated artifact rejection for MEG and EEG data. NeuroImage 159, 417–429.

Jas, M., Larson, E., Engemann, D.A., Leppäkangas, J., Taulu, S., Hämäläinen, M., Gramfort, A., 2018. A reproducible MEG/EEG group study with the MNE software: Recommendations, quality assessments, and good practices. Frontiers in Neuroscience 12, 1–18.

Jezzard, P., Balaban, R.S., 1995. Correction for geometric distortion in echo planar images from B0 field variations. Magnetic Resonance in Medicine 34, 65–73.

Jungl, T.p., Humphriesl, C., Lee, T.w., Makeig, S., 1998. Extended ICA Removes Artifacts from Electroencephalographic Recordings. Advances in Neural Information Processing Systems 10, 894–900.

Khan, S., Hashmi, J.A., Mamashli, F., Michmizos, K., Kitzbichler, M.G., Bharadwaj, H., Bekhti, Y., Ganesan, S., Garel, K.L.A., Whitfield-Gabrieli, S., Gollub, R.L., Kong, J., Vaina, L.M., Rana, K.D., Stufflebeam, S.M., Hämäläinen, M.S., Kenet, T., 2018. Maturation trajectories of cortical resting-state networks depend on the mediating frequency band. NeuroImage 174, 57–68.

Koutsouleris, N., Davatzikos, C., Borgwardt, S., Gaser, C., Bottlender, R., Frodl, T., Falkai, P., Riecher-Rossler, A., Moller, H.J., Reiser, M., Pantelis, C., Meisenzahl, E., 2014. Accelerated Brain Aging in Schizophrenia and Beyond: A Neuroanatomical Marker of Psychiatric Disorders. Schizophrenia Bulletin 40, 1140–1153.

Liem, F., Varoquaux, G., Kynast, J., Beyer, F., Kharabian Masouleh, S., Huntenburg, J.M., Lampe, L., Rahim, M., Abraham, A., Craddock, R.C., Riedel-Heller, S., Luck, T., Loeffler, M., Schroeter, M.L., Witte, A.V., Villringer, A., Margulies, D.S., 2017. Predicting brain-age from multimodal imaging data captures cognitive impairment. NeuroImage 148, 179–188.

Loeffler, M., Engel, C., Ahnert, P., Alfermann, D., Arelin, K., Baber, R., Beutner, F., Binder, H., Brähler, E., Burkhardt, R., Ceglarek, U., Enzenbach, C., Fuchs, M., Glaesmer, H., Girlich, F., Hagendorff, A., Häntzsch, M., Hegerl, U., Henger, S., Hensch, T., Hinz, A., Holzendorf, V., Husser, D., Kersting, A., Kiel, A., Kirsten, T., Kratzsch, J., Krohn, K., Luck, T., Melzer, S., Netto, J., Nüchter, M., Raschpichler, M., Rauscher, F.G., Riedel-Heller, S.G., Sander, C., Scholz, M., Schönknecht, P., Schroeter, M.L., Simon, J.C., Speer, R., Stäker, J., Stein, R., Stöbel-Richter, Y., Stumvoll, M., Tarnok, A., Teren, A., Teupser, D., Then, F.S., Tönjes, A., Treudler, R., Villringer, A., Weissgerber, A., Wiedemann, P., Zachariae, S., Wirkner, K., Thiery, J., 2015. The LIFE-Adult-Study: Objectives and design of a population-based cohort study with 10,000 deeply phenotyped adults in Germany. BMC Public Health 15, 1–14.

López-Otín, C., Blasco, M.A., Partridge, L., Serrano, M., Kroemer, G., 2013. The Hallmarks of Aging. Cell 153, 1194–1217.

Meda, S.A., Giuliani, N.R., Calhoun, V.D., Jagannathan, K., Schretlen, D.J., Pulver, A., Cascella, N., Keshavan, M., Kates, W., Buchanan, R., Sharma, T., Pearlson, G.D., 2008. A large scale (N=400) investigation of gray matter differences in schizophrenia using optimized voxel-based morphometry. Schizophrenia Research 101, 95–105.

Mensch, A., Mairal, J., Thirion, B., Varoquaux, G., 2016. Dictionary Learning for Massive Matrix Factorization, in: Balcan, M.F., Weinberger, K.Q. (Eds.), Proceedings of The 33rd International Conference on Machine Learning, PMLR, New York, New York, USA. pp. 1737–1746.

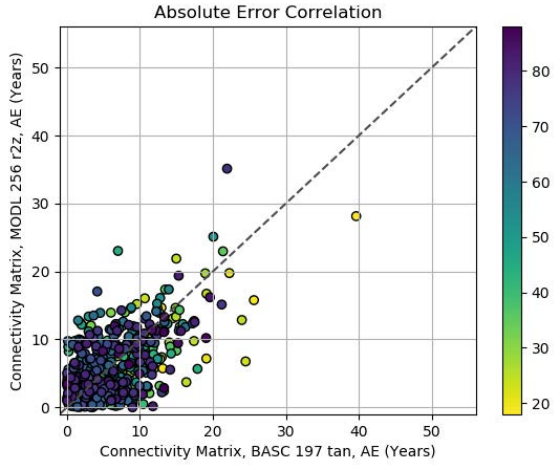
Niso, G., Rogers, C., Moreau, J.T., Chen, L.Y., Madjar, C., Das, S., Bock, E., Tadel, F., Evans, A.C., Jolicoeur, P., Baillet, S., 2016.

OMEGA: The Open MEG Archive. NeuroImage 124, 1182–1187.
Nooner, K.B., Colcombe, S.J., Tobe, R.H., Mennes, M., Benedict, M.M., Moreno, A.L., Panek, L.J., Brown, S., Zavitz Stephen, T.T., Li, Q., Sikka, S., Gutman, D., Bangaru, S., Schlachter, R.T., Anwar, S.M., Hinz, C.M., Kaplan, M.S., Rachlin, A.B., Adelsberg, S., Cheung, B., Khanuja, R., Yan, C., Courtney, C.C., King, M., Wood, D., Cox, C.L., Kelly, A.M., Petkova, E., Reiss, P.T., Duan, N., Thomsen, D., Biswal, B., Coffey, B., Hoptman, M.J., Javitt, D.C., Pomara, N., Sidtis, J.J., Koplewicz, H.S., Castellanos, F.X., Leventhal, B.L., Milham, M.P., 2012. The NKI-Rockland sample: A model for accelerating the pace of discovery science in psychiatry. Frontiers in Neuroscience 6, 1–11.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Percival, D.B., Walden, A.T., 1993. Spectral Analysis for Physical Applications. Cambridge University Press.
Pfefferbaum, A., Mathalon, D.H., Sullivan, E.V., Rawles, J.M., Zipursky, R.B., Lim, K.O., 1994. A Quantitative Magnetic Resonance Imaging Study of Changes in Brain Morphology From Infancy to Late Adulthood. Archives of Neurology 51, 874–887.
Poldrack, R.A., Nichols, T., Mumford, J., 2011. Handbook of Functional MRI Data Analysis. Cambridge University Press, Cambridge.
Price, D., Tyler, L.K., Neto Henriques, R., Campbell, K.L., Williams, N., Treder, M., Taylor, J.R., Henson, R.N.A., 2017. Age-related delay in visual and auditory evoked responses is mediated by white- and grey-matter differences. Nature Communications 8, 15671.
Rahim, M., Thirion, B., Comtat, C., Varoquaux, G., 2016. Transmodal Learning of Functional Networks for Alzheimer's Disease Prediction. IEEE Journal on Selected Topics in Signal Processing 10, 1204–1213.
Shafto, M.A., Tyler, L.K., Dixon, M., Taylor, J.R., Rowe, J.B., Cusack, R., Calder, A.J., Marslen-Wilson, W.D., Duncan, J., Dalgleish, T., Henson, R.N., Brayne, C., Matthews, F.E., 2014. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. BMC Neurology 14, 1–25.
Storsve, A.B., Fjell, A.M., Tamnes, C.K., Westlye, L.T., Overbye, K., Aasland, H.W., Walhovd, K.B., 2014. Differential Longitudinal Changes in Cortical Thickness, Surface Area and Volume across the Adult Life Span: Regions of Accelerating and Decelerating Change. Journal of Neuroscience 34, 8488–8498.
Taulu, S., Kajola, M., 2005. Presentation of electromagnetic multichannel data: The signal space separation method. Journal of Applied Physics 97.
Taylor, J.R., Williams, N., Cusack, R., Auer, T., Shafto, M.A., Dixon, M., Tyler, L.K., Cam-CAN, Henson, R.N., 2017. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. NeuroImage 144, 262–269.
Tipping, M.E., 2001. Sparse Bayesian Learning and the Relevance Vector Machine. J. Mach. Learn. Res. 1, 211–244.
Uusitalo, M.A., Ilmoniemi, R.J., 1997. Signal-space projection method for separating MEG or EEG into components. Medical and Biological Engineering and Computing 35, 135–140.
Varoquaux, G., Baronnet, F., Kleinschmidt, A., Fillard, P., Thirion, B., 2010. Detection of Brain Functional-Connectivity Difference in Post-stroke Patients Using Group-Level Covariance Modeling, in: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2010, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 200–208.
Varoquaux, G., Raamana, P.R., Engemann, D.A., Hoyos-Idrobo, A., Schwartz, Y., Thirion, B., 2017. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. NeuroImage 145, 166–179.
Vlahou, E.L., Thurm, F., Kolassa, I.T., Schlee, W., 2014. Resting-state slow wave power, healthy aging and cognitive performance. Scientific Reports 4, 33–36.
Vos, T., Flaxman, A.D., Naghavi, M., Lozano, R., Michaud, C., Ezzati, M., Shibuya, K., Salomon, J.A., 2012. Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 19902010: a systematic analysis for the Global Burden of Disease Study 2010. The Lancet 380, 2163–2196.
Welch, P., 1967. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. IEEE Transactions on Audio and Electroacoustics 15, 70–73.
Wolpert D., H., 1992. Stacked generalization. Neural Networks 5, 241–259.
Yarkoni, T., Westfall, J., 2017. Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. Perspectives on Psychological Science 12, 1100–1122.

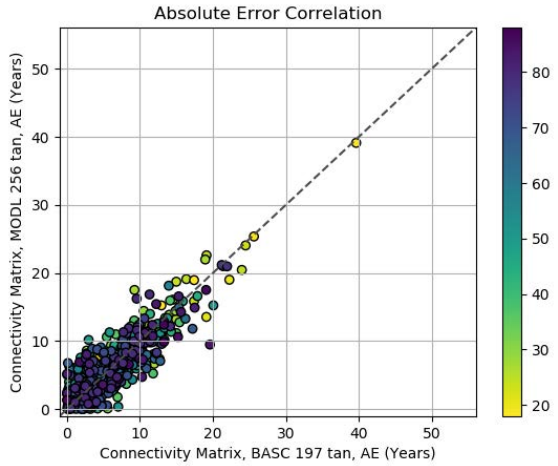
Appendix A. Supplementary results

Modality	MAE	STD	R2
Cortical Surface Area	10.39	1.00	0.48
Cortical Thickness	7.31	0.64	0.72
Subcortical Volumes	6.76	0.60	0.76
BASC 197 tan	5.85	0.51	0.82
BASC 197 r2z	6.74	0.40	0.76
MODL 256 tan	5.90	0.46	0.82
MODL 256 r2z	5.82	0.40	0.83
MEG	9.22	0.56	0.58
MEG, CSA Stacked	7.64	0.84	0.71
MEG, CT Stacked	6.45	0.57	0.80
MEG, SV Stacked	5.89	0.56	0.83
MEG, BASC 197 tan Stacked	5.59	0.79	0.84
MEG, MODL 256 r2z Stacked	5.72	0.34	0.84
MRI Stacked	6.16	0.49	0.82
fMRI Stacked	5.94	0.63	0.82
MRI, fMRI Stacked	5.09	0.38	0.87
MEG, MRI Stacked	5.43	0.50	0.86
MEG, fMRI Stacked	5.25	0.65	0.86
MEG, MRI, fMRI Stacked	4.75	0.46	0.89

Table A.1: Performance metrics for different models. Mean absolute error, R^2 score were measured using 10-fold cross-validation. STD is a standard deviation of MAE. `CSA`, `CT`, and `SV` are abbreviations of cortical surface area, cortical thickness, and subcortical volumes correspondingly. FMRI features estimated in the tangent space are denoted using the `tan` abbreviation, while `r2z` denotes fMRI features calculated using Fisher's r-to-z transformation. Multi-source models (`Stacked`) perform better than single-source models, stacking with MEG data improves the age prediction accuracy. The highest accuracy was achieved by the model combining MEG, structural and functional MRI, namely the `MEG, MRI, fMRI Stacked` model.



(a)

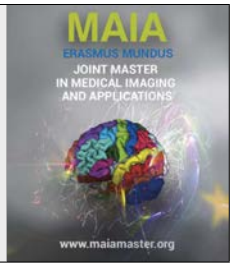


(b)

Figure A.8: Correlation of absolute error between the models trained on fMRI features. The color bar was added to indicate the age of a subject. `MODL 256 r2z` is a connectivity matrix estimated for the MODL parcellation atlas and further processed using Fisher's r-to-z transformation, `MODL 256 tan` is a connectivity matrix estimated for the MODL parcellation atlas in the tangent space. `BASC 197 tan` is a connectivity matrix estimated for the BASC parcellation atlas in the tangent space. Correlation between fMRI features MODL 256 r2z and BASC 197 tan are lower than the one of the features estimated in the tangent space for both atlases.

Medical Imaging and Applications

Master Thesis, June 2019



Specified Metal Artifact Reduction (MAR) on CT-scan for dosimetry accuracy in I-125 prostate brachytherapy

Antoine Merlet, Alain Lalande, Gilles Crhange

ImVia, Universit Bourgogne-Franche Comt, Dijon, France

Abstract

Prostate cancer is the most common and the second deadliest cancer for males. Low-dose brachytherapy is based on the implant of radioactive seeds directly in the prostate, and has become the technique of choice to cure prostate cancer. In order to maximize the dose delivered to the cancerous area while minimizing the radiation effects on healthy organs, both treatment planning and follow-up are necessary. Computed Tomography (CT) is the standard imaging modality to perform radiation therapy dosimetry, but the implanted seeds are made of metallic coating, inducing severe artifacts in CT images. While many Metal Artifact Reduction (MAR) methods have been proposed and implemented in the past four decades, the focus has been on artifacts caused by orthopaedic, dental and hip implants. This paper presents a MAR post-processing algorithm based on prostate CT images. The effect of the proposed correction is evaluated on 33 CT-scans of post-implant patients, and the changes in TG-43 compliant dose distribution calculation are discussed. Results show a change in the absolute number of seeds detected after correction (± 2.9 seeds on average), inducing minor changes in dose calculation. There were no significant changes in D_{90}, V_{100} and V_{150} when seed detection is consistent.

Keywords: Brachytherapy, Prostate cancer, Dosimetry, Metal artifact reductiona, Computed tomography

1. Introduction

Prostate cancer is the most common type of cancer diagnosed for males (behind skin cancers), and the second leading cause of cancer related deaths. The American Cancer Society (ACS) estimates that 31 620 prostate-cancer related deaths are to be expected in 2019, from 174 650 newly diagnosed cases (Siegel et al., 2019). Prostate cancer can be detected using a wide variety of techniques: Prostate Specific antigens (PSA), Ultrasound (US), Magnetic Resonance Imaging (MRI) or digital rectal examination (non-exhaustive list) (Rozet et al., 2018). The most straightforward way to treat prostate cancer is radical prostatectomy, but this operation induces irreversible and undesirable side effects for the patient, such as sterility, incontinence or erectile dysfunction. Research has been done on alternative treatment methods in order to improve treated patients' well-being, and several are now available. Nowadays, according to the diagnosis, it is often possible for the patient to choose among different available techniques based on their respective side effects. This leads to a common use of radiation therapy to cure prostate cancer, as its downsides are often considered to be minor. This method aims to destroy cancerous cells by using radiations, delivered using either External Beam Radiation Therapy (EBRT) or BrachyTherapy (BT). Low-Dose-Rate BrachyTherapy (LDR-BT) is a form of radiation therapy, for which small metallic seeds containing a radioactive source are inserted locally inside the patients' body in order to deliver the radiations. With this procedure, about 50 to 100 Iodine-125 (I-125) seeds are permanently introduced into the prostate. BT allows for localized treatment of low to middle grade cancer. Due to the radioactive nature of the treatment and the close proximity of the prostate to other tissues such as bladder and rectum, it is of paramount importance for the delivered radioactive dose to be maximized on the cancerous cells, while being minimized for the surrounding healthy tissues (Créhange et al., 2017).

This brings a need for accurate dose calculation

Figure 1: Dosimetry analysis on post-implant CT-scans. The prostate and rectum are outlined in red and lemon green, respectively. In magenta, green and blue are represented the cumulative isodose line for 240 Gy, 180 Gy and 120 Gy respectively.

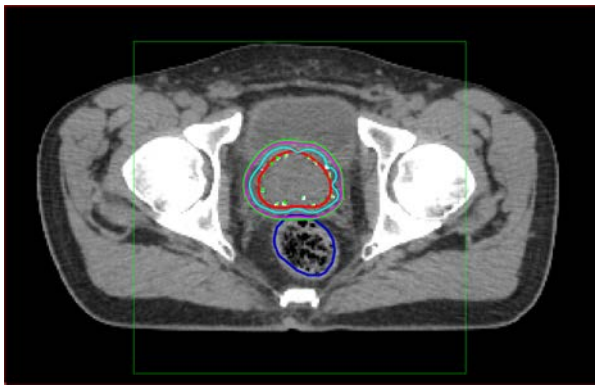
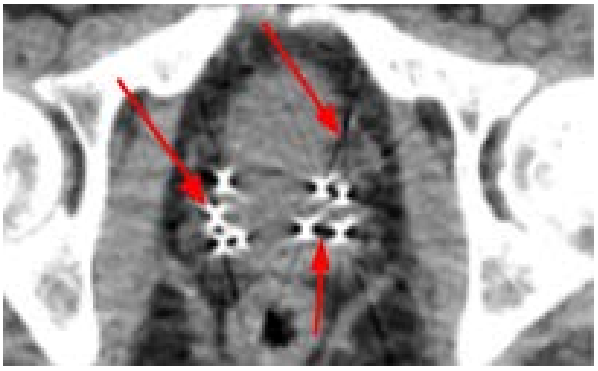


Figure 2: CT slice of post-implant patient zoomed at the level the prostate. Are shown by red arrows the two types of artifacts: cupping (bright and dark regions around seeds) and streaking (dark bands).



(also called dosimetry). According to the American Brachytherapy Society recommendations (Nag et al., 2000; Rivard et al., 2007), dosimetry study should be performed after the seed implantation. The protocol for brachytherapy dose calculations has been extensively described by the American Association of Physicists in Medicine (AAPM) Task Group No. 43 over several reports and updates (Rivard et al., 2004). Updates are being published, according to the evolution of technologies in brachytherapy. It aims to define meaningful metrics for clinical evaluation, such as the dose covering 90% of the prostate (D_{90}), the fractional volume of the prostate receiving 100% of the prescribed dose (V_{100}), or the urethral and rectal doses received (see Figure 1). There is now a standard procedure to input CT data of post-implant patient into commercially available software in order to obtain dosimetry data.

However, it is known that in the presence of metallic objects such as prosthesis, dental implants or brachytherapy sedds, X-ray and CT images are subject to corruption by artifacts (Barrett and Keat, 2004; Boas and Fleischmann, 2012). These metal artifacts are linked to several mechanisms. The most impacting one is called beam hardening. It occurs when a X-ray beam encounters a high density object, where only high energy photons are able to go through the object, resulting in ”hardening” the beam (i.e. increasing its mean energy). This results in two types of artifacts, called streaking (dark and bright bands) and cupping (dark regions) artifacts, as shown in Figure 2. Correction of such artifacts is necessary, as their presence may affect any kind of segmentation based methods, clinical diagnosis or even visual evaluation of the images. The influence of Metal Artifact Reduction (MAR) algorithms on CT-scans have been shown to bring improvements, for both dosimetry analysis and ease of organ recognition (Bazalova et al., 2007; Giantsoudi et al., 2017).

While it is now common for CT-scanner manufacturers to provide built-in MAR techniques with their products, the proposed correction can be improved for specific cases, and beam hardening correction remains a challenging task (Andersson et al., 2015; Huang et al., 2015; Bolstad et al., 2018). Moreover, several studies consider MAR caused by metal object of different sizes, but very few consider a large quantity of metallic elements (and therefore many possible metal artifacts sources). Regarding prostate radioactive seed implant, very small objects in high quantity have to be considered separately, making MAR a challenging task when applied to post-implant CT-scans since conventional MAR methods are not well-adapted (Karimi et al., 2012).

However, it has been shown that the use of non-seed specified MAR algorithms yields a positive impact on post-implant dosimetry (Andersson et al., 2015). The positive impact of two different scanner specific MAR algorithms applied to lung and prostate brachytherapy respectively has been demonstrated (Yang et al., 2015; Shiraishi et al., 2016). Many MAR methods are based on sinograms, which are commonly stored using proprietary format, and therefore often not accessible (Basran et al., 2011). Finally, the validation of dosimetry accuracy improvements due to MAR is often performed on phantom data, as patient ground-truth data does not usually exist.

To the best of the our knowledge, no seed-specific MAR method have been proposed while being accompanied by the dosimetry changes induced by the correction. This works aims to: *i)* develop a MAR method specific for brachytherapy seed without access to raw sinogram data, *ii)* evaluate the dose calculations after MAR using a commercial TPS.

2. State of the art

While introduced several decades ago, the formalism for dose calculation of brachytherapy radionuclide

sources proposed by the Task Group No. 43 (TG-43) of the American Association of Physicists in Medicine (AAPM) is still today recognized as a worldwide standard (Nath, 1995). The suggested protocols have been updated over time, with respect to the technological advances applied to brachytherapy (Rivard et al., 2004; Rivard et al., 2007). In 2012, the TG-186 released guidelines for early adopter of Model-Based Dose Calculation Algorithms (MBDCAs) (Beaulieu et al., 2012). These MBDCAs are a novelty in the field, and it is necessary to correlate their results with the TG-43 formalism. In this section is first presented a brief overview of the TG-43 formalism and its major implications regarding clinical dose calculations. Secondly, the TG-186 formalism for MBDCAs will be presented with its improvements over the TG-43 formalism. Finally, a brief review of common MAR concepts is given according to their use in commercial CT-scanners, with specific state-of-the-art for BT seeds.

2.1. TG-43 formalism

According to the general dosimetry formalism proposed by the TG-43 (TG-43, TG-43U1, TG-43U1S1 and TG-43U1S2 reports), two different dose calculation approaches are defined: 1D (point source) dose calculation and 2D (cylindrically symmetric line source) dose calculation. They describe the dose distribution around a single source centrally positioned in a spherical water phantom.

This formalism give consideration for the photon fluence around the source in scattering medium (i.e. water), in opposition with older (pre TG-43 formalism) protocols considering dose distribution in free space. This transition is done by performing dose distribution measurement in a water equivalent medium. The estimation of the dosimetry of a seed model is generally performed using a Lithium Fluoride (LiF) ThermoLuminescent Dosimeter (TLD), or by one-time source-model specific Monte Carlo (MC) simulation (Briesmeister et al., 2000; Kawrakow, 2001). Aside from minor parameter value changes since the first TG-43 report, the 1D and 2D dose-rate equations remain unchanged (Eq. 1 and Eq. 2 respectively).

$$\dot{D}(r) = S_K \cdot \Lambda \cdot \frac{G_L(r,\theta_0)}{G_L(r_0,\theta_0)} \cdot g_L(r) \cdot \phi_{an} \qquad (1)$$

$$\dot{D}(r,\theta) = S_K \cdot \Lambda \cdot \frac{G_L(r,\theta)}{G_L(r_0,\theta_0)} \cdot g_L(r) \cdot F(r,\theta) \qquad (2)$$

The equations are generalized for all type of sources (i.e. encapsulated radioactive materials). In this work, the source is specified as a seed (i.e. cylindrically symmetric brachytherapy source of effective length less or equal to 0.5 cm).

While giving the reader a full understanding of these equations and their parameters are out of the scope of this work, it is essential to understand their principle, assumptions and limitations. Therefore, only a simplified description of the parameters is given. For a more detailed description of the parameters and functions, please refer to the latest TG-43 updates (TG-43U1 and TG-43U1S1). See Appendix 1 for an exact definition of the terms used below.

For a given point of interest, with X being either 'P' for point-source (1D) or 'L' for line-source (2D) depending on the approach considered in the geometry and radial dose functions, and with the assumption of cylindrically symmetric source in a water transport medium:

- r : Distance of the point of interest to the seed center (cm);
- θ : Polar angle of the point of interest w.r.t the seed longitudinal axis;
- r_0 : Reference distance ($1\,cm$ according to this standardize protocol);
- θ_0 : Reference angle defining the transverse plane of the seed ($90°$ or $\pi/2$ radians);
- S_K : Air-kerma strength (kerma: kinetic energy released per unit mass) of the seed in $\mu Gy\, m^2\, h^{-1}$ (or more commonly, U). This value differ according to the seed model;
- Λ : Dose-rate constant in water. Linked to S_K (ratio);
- $G_X()$: Geometry function used to interpolate and/or extrapolate data tabulated at discrete points;
- $g_X()$: Radial dose function, describing the dose fall-off on the transverse plane due to photon attenuation and scattering;
- $F()$: 2D anisotropy function, representing the variation in dose according to the polar angle (w.r.t the transverse plane);
- $\phi_{an}()$: 1D anisotropy function.

The main conceptual difference between the two approaches is the shape-model used. While Eq. 2 takes into account both the seed location and orientation for dose distribution calculation, Eq. 1 is an approximation of the previous equation in which the seed is considered as an isotropic point-source. Therefore, the 1D formalism is taking into account, on a geometrical level, only the radial distance of a point of interest from the source center as parameter.

Overall, the TG-43 formalism allows for a common ground between institutions regarding dose calculation practices, Treatment Planning System (TPS) providers and seed manufacturers. The combined use of the previous mathematical models and Monte Carlo simulations enables fast dosimetry in clinical environment, and the TG-43 formalism is respected in the vast majority of TPSs.

2.2. TG-186: Model-Based Dose Calculation Algorithms

The limitations induced by the TG-43 formalism have been quantified (Rivard et al., 2009), and it has been shown a reduction of 2% to 10% regarding the estimation of the D_{90} parameter. In report describes several general-purpose methods (i.e. non organ specified radiation therapy) considering a nonwater transport medium for the photons as well as inter-seed attenuation, which have been shown to more accurately estimate the dose delivered to the patient than with the TG-43 formalism. With time, these methods are being specified for their application, such as prostate brachytherapy.

In 2012, the AAPM TG-186 released recommendations regarding the early use of Model-Based Dose Calculation Algorithms (MBDCAs). Model-based approaches either explicitly simulate the transport of radiation in the actual media or employ multiple dimensional scatter integration techniques to account for the dependence of scatter dose on the 3D geometry (Beaulieu et al., 2012). MBDCAs have been classified in three groups according to their approach: Collapsed-cone method (Ahnesjö, 1989), Monte Carlo simulations solving the Linear Boltzmann Transport Equation (LBTE) (Williamson, 1987), and Grid-Based Boltzmann Equation Solvers (GBBS) (Zhou and Inanc, 2002).

As of 2017, there are three commercial TPS using MBDCAs, with none for LDR (Rivard et al., 2017). The absence of commercial MBDCAs for LDR can be attributed to the challenge of accurately quantifying the absorbed dose in tissues (Rivard et al., 2009). The absorbed dose is dependent on the transport medium composition and density, which can be estimated using CT imaging. However, it has been shown that the presence of artifacts in CT images can lead to discrepancies in dose calculation as the tissue density estimation (Hounsfield Units, HU, that is a relative tissue density) might be corrupted (Bazalova et al., 2007). The TG-186 stated that the accuracy of brachytherapy dose calculations could be improved by the development of CT-scan metal artifact correction, especially in seed vicinity.

2.3. Metal Artifact Reduction

In their work, Gjesteby et al., 2016 gave an overview of the evolution of MAR in the past decades. It is clearly shown a high interest for MAR, especially in the last 10 years: 20 MAR related publications in 2009, for 95 in 2015. According to the literature, MAR can be performed on different levels, which are presented below together with the approach taken by four CT-scanner manufacturers.

A first common approach is to correct data in the sinogram space (called Projection Completion methods), as the CT values within the metal trace are often missing or corrupted. This correction can be performed either by direct interpolation from neighboring values in the sinogram space (Veldkamp et al., 2010), by data normalization (NMAR, Meyer et al., 2010, Siemens company), or by iterative forward reprojection of a prior image. The iterative forward reprojection approach is followed by:

- Single-Energy Metal Artifact Reduction (SEMAR) algorithm, where a prior image is obtained from first-pass metal artifact reduction by linear interpolation in the sinogram, and is then used for forward reprojection (Chang et al., 2012), Toshiba company;
- Orthopedic-MAR (O-MAR) algorithm, where an iterative framework is used, in which the output corrected input is subtracted from the original input image, and then becomes the new input image of the iterative framework (Healthcare, 2012), Philips company.

Once the correction is performed in the sinogram space, Filtered-Back Projection (FBP) is often used to project the sinogram into the image space. However, as the FBP assumes that the projection data is complete and perfect, projection completion methods might induce secondary artifacts when the correction is not accurate, such as blurring around the metal object (due to data interpolation).

A second approach for MAR is the Iterative Reconstruction, where the FPB is replaced with objective function optimization algorithms, such as Maximum-Likelihood (ML), Expectation-Maximization (EM) or Algebraic Reconstruction Technique (ART) (Mouton et al., 2013). While Iterative Reconstruction methods generally produce a better final result than sinogram completion methods, their iterative schemes often have a high computational cost, limiting their clinical use.

Thirdly, other methods try to take into account the advantages of the two previous MAR approaches. This is the case for the hybrid Iterative Metal Artifact Reduction (IMAR) proposed by Siemens company (Axente et al., 2015). This algorithm iteratively combines two of their previously developed methods: Normalized MAR (NMAR, projection completion) and Frequency Split MAR (FSMAR, iterative reconstruction method) (Meyer et al., 2010; Meyer et al., 2012). General Electrics company implemented MAR is their Gemstone Spectral Imaging (GSI) Dual-Energy CT (DECT) using fast kV-switch to obtain a Virtual Monochromatic Spectral (VMS) image (Lee et al., 2012, Pessis et al., 2013). Low-energy CT imaging (e.g. 80 keV) enable high tissue contrast but is subject to high beam hardening effect, while higher energy imaging (e.g. 140 keV) allows for better differentiation of the metal-bone interfaces, while having a poor tissue contrast. By trying to combine the advantages of each energy imaging while minimizing their drawbacks, the produced VMS has a

high tissue contrast, and the usually photon-starved regions are corrected.

The methods mentioned above all have a common component: the availability of the raw data (sinogram). While it might be easily obtainable for constructors and/or in some research environments, the raw data is commonly stored in proprietary format, making it hardly obtainable in a clinical environment, where only post-FBP (reconstructed) CT-scans are available. MAR methods performed in the image space are often call post-processing methods. Such methods generally yields worse results than all previous methods, as the corruption in the metal trace is propagated in the image space by the FPB. While it is possible to reconstruct a virtual sinogram using the inverse Radon Transform (iRT), the obtained sinogram will contain more corruption than the true original sinogram. Post-processing MAR is under-developed compared to all the previously mentioned methods, and the usual metal trace segmentation has to be replaced by accurate seed detection in the image space in order to perform MAR.

Overall, almost all MAR algorithm proposed in the literature are based on sinogram correction and/or are applied to specific metal implants, such as orthopedic or dental filling (Mouton et al., 2013; Huang et al., 2015; Andersson et al., 2015; Gjesteby et al., 2016; Bolstad et al., 2018). Few considerations are given to MAR applied to artifacts generated by the presence of BT seeds.

The first work found in literature applied to prostate BT seeds proposed a HU value thresholding method in the image space to generate a metal only image (Takahashi et al., 2006). Both the uncorrected image and the metal only image are then converted into projection space using iRT, and the metal sinogram is substracted from the uncorrected sinogram. The resultant is then converted back to the image space using the Radon transform. This simple approach allowed for seed identification, but only minor improvements over streaking artifacts were shown. They correlated the reduction of artifacts with the reduction of the Standard Deviation (SD) in the Region Of Interest (ROI).

In their work, Xu et al. (2011) present a raw sinogram based MAR for BT seeds, where the metal trace detection was performed using the Steger method (Steger, 1998). Missing data were generated using linear interpolation using adjacent projection, and the corrected sinogram was converted to the image space using FBP. They showed an increase only on the D_{90} parameter, with plus 12% after correction with their method (one case study).

Finally, a single study proposes a post-processing MAR method applied to prostate BT (Basran et al., 2011). Their approach takes advantage of the information contained in the adjacent CT slices by using a 3-D median filter. First, the dark bands are identified using a CT value threshold combined with a threshold on the standard deviation of the adjacent voxels in the azimuthal direction. Then, the values of the voxels within the dark bands are replaced with the median of the ROI. Finally, an evenly-weighted 3-D median filter is applied on the image (at the exception of seeds). Its size is so that it covers a $5 \times 5 \times 5\, mm^3$ area. While this type of processing efficiently reduces the amount of visible artifacts, it also reduces contrast in regions containing bones and/or areas containing air. The influence of this correction on dose calculation was not given, but the authors mentioned that it is expected to have no impact (in the case of TG-43 compliant dose calculations).

In this work is presented a new post-processing MAR method when raw sinogram data is not available. It involves *(1)* seed area extraction; *(2)* iterative intensity-based clustering of voxels in seed proximity; *(3)* linear interpolation of corrupted data; *(4)* dosimetry analysis of the corrected image according, to the TG-43 formalism.

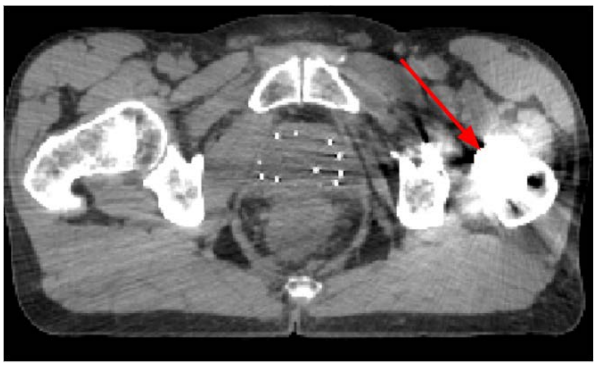
3. Material and methods

3.1. Experimental data

The dataset used for this work is composed of prostate CT-scans from 33 post-implant patients (primary and salvage brachytherapy). Data were obtained from a UNCANCER-affiliated center (CGFL, Dijon, France) from clinical cases were the patient was treated for prostate cancer, and where post-implant dosimetry analysis was performed by a physician. All patient CT-scans have a voxel resolution of $1.2695 \times 1.2695\, mm^2$, with a slice thickness of $1.25\, mm$. All CT-scans values were converted to Hounsfield Units (HU).

The number of implanted seeds in a single patient is varying from 17 to 97 (60 ± 22 seeds on average). The only selection criteria was the availability of the prostate segmentation performed by an expert. Therefore, there is a high variety in patient data. Some cases may contain metallic object(s) other than the seeds (such as hip prosthesis), leading to an external (w.r.t prostate area) source of metal artifacts, as shown in Figure 3.

Figure 3: CT slice of a patient having a metallic hip prosthesis (red arrow), being a source metal artifacts exterior to the prostate.



All patients were treated using the selectSeed I-125 cylindrically symmetric source model (Karaiskos et al., 2001), of size $L = 4.5\,mm$ and radius $r = 0.4\,mm$. The prescribed peripheral doses for the 33 cases were ranging from 90 Gy to 160 Gy, with activities between 0.372 U to 0.789 U.

Seed locations and dosimetry data were extracted for each case from the VariSeed treatment planning system (VariSeed 9.0, Varian Medical Systems, Palo Alto, CA). Seed coordinates were converted from world space to image space using the voxel resolution information. Since the seed positions extracted from VariSeed have been used in the clinical assessment of patient dose distribution, one could assume that their number and location detection as ground-truth data. The dosimetry analysis performed by the used TPS is in compliance with the TG-43 point-source formalism (i.e. 1D approach).

All data were stored using the DICOM format, and all figures showing CT data in HU have the window and level display parameter of 400 and 40 respectively (unless stated otherwise).

3.2. Metal Artifact Reduction framework

The framework for MAR has been designed according to three main ideas, the first one being that artifacts produced by the seeds are always connected to the seed itself and/or to the overestimation of the seed area due to the Partial Volume Effect (PVE). The second point is that, for any given slice containing a seed, the information contained in the adjacent axial slices can be useful for both metal artifact detection and corrupted data interpolation. Thirdly, the unavailability of ground-truth data (metal artifact free post-implant images) for a given patient limits the use of supervised methods.

In the early stages of this work, we tried to solve the MAR problem using machine learning methods. A biclass balanced dataset of 10 000 3-D patches was created, where the first class was composed of artifact-free patches, and the second class was made of patches containing metal artifacts. 10% of patches were kept aside for the final testing of the proposed framework. Patches where extracted from manual selection (based on visual evaluation) of specific voxels locations from all 33 patient cases. Therefore, as each patch is defined by a central voxel, on could easily change the patch size. Extraction was performed so that there was a high intra-class variety regarding patches content. One example is, for the second class, the selection of patches close to seeds, bones or prostate contour. The approach taken was the use of AutoEncoders (AE) for feature extraction combined with Support Vector Machine (SVM) for a classification task using the latent space representation of the data in the AE as input (Baldi, 2012; Cortes and Vapnik, 1995).

Several AE architectures have been designed, with different number of filters, depth and input shape. Two training configurations have been used: training of a single AE on the full dataset regardless of the class, and training of one AE per class (using the same architecture for both classes). In the first case, the aim is to use the features extracted at the end of the encoding part of the AE as the input of a SVM for direct classification. In the second case, we wanted to build two distinct AE, each able to accurately reconstruct one and only one class. After training, both AE were used in order to produce a latent representation of the test patches. This method was expected to produce aberrant results when a patch of a given class was reconstructed using the inappropriate AE, therefore allowing the discrimination of the two produced representations for the classification task.

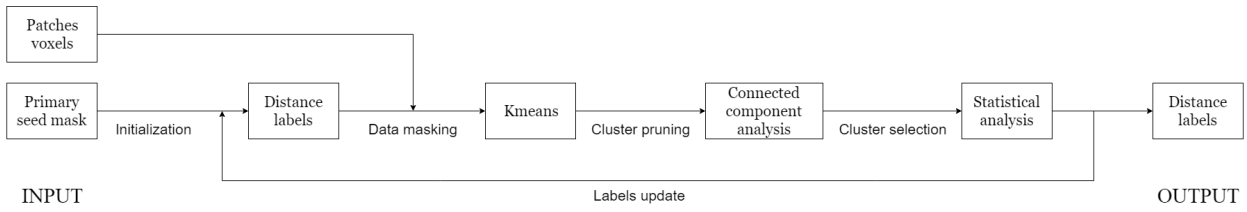
The two best models for the reconstruction task were a 3-D AE and a 2.5-D AE, for all training configurations. The 3-D AE was trained with the extracted 3-D patches. The 2.5-D approach was trained for the reconstruction of a patch central slice using information contained not only in the considered patch slice, but also in the two adjacent axial slices (where each slice was considered as a distinct channel).

While the proposed architectures and training configurations could accurately reconstruct the input data, the use of the representation of the data in the latent space as the input of SVMs was inconclusive for all proposed configurations, with the best area under the curve being 0.81. In section 5 are given our thoughts regarding the inefficiency of this method.

In the light of the results obtained with the first developed framework, we tried a different method, being a patient-specific approach based on intensity clustering of seed-centered patches. The number of patches extracted from a given patient is equal to the number of detected seeds in the considered patient. In the remaining of this work, the word 'patches' will refer to all the extracted seed-centered patches for a given patient. The radius of the patch used is 10 x 10 x 3 pixels for all patients. The choice of patch with a large length in the coronal and sagittal directions is motivated by the fact that metal artifacts originating from a specific seed mostly produce visible effect at a close to medium distance from considered seed. Moreover, when the patch size is increased (in the previously mentioned directions), the ratio of uncorrupted voxels over the total number of voxels will generally increase, thus easing the identification of 'outliers' (metal artifacts) using clustering methods. Furthermore, due to the properties inherent to 3D CT-scan construction from axial slices acquisition, no streak will propagate in the axial direction, hence the choice of smaller radius in this direction. For clustering purposes, the values contained in a patient patches can be seen as a 1-D vector, initially of size 21 x 21 x 7 x number of seeds.

Due to proximity, a seeds (called secondary) might appear in a patch centered on another seed (called primary). In this case, the HU values of the all secondary

Figure 4: Iterative clustering framework workflow.



seeds (w.r.t primary seeds) are excluded from further calculation. For each patch, only the main seed is considered and segmented, resulting in a mask hereinafter called primary seed mask. Any secondary seed voxels will be excluded from further calculation (see Figure 5). However, duplicated data are kept, i.e. any patch may share voxels with any other patch. Since a patch-based centered on the seeds is taken, and knowing that only seeds inside the prostate are considered, one can be sure that no patch will be centered on the pelvic bone (having high HU values). Therefore, we chose to perform seed segmentation by performing a single K-means clustering with k = 2 on all patches data (Lloyd, 1982). Due to its working principle, this method will consistently create one cluster containing the highest HU values being the seeds, and voxels with lower HU values, such as tissue form the second cluster. It is noteworthy that voxels having medium HU value induced to Partial Volume Effect (PVE) are consistently clustered with tissues. While these voxels values are high compared to tissue values, they are relatively low compared to seed HU values, hence their clustering with tissue data.

The obtained primary seed mask is used for the initialization of distance map D, which is then used in iterative framework where remaining voxels are iteratively clustered according to their intensities (see Figure 4). The first cluster is therefore the seeds themselves, and they take a label noted L in the distance map D, noted $D_{L=1}$, while voxels to be clustered have a different label, noted $D_{L=0}$. On each subsequent iteration, the remaining data of all patches, noted $P(D_{L=0})$ is clustered using K-means. In this work are consider 8 distinct clusters ($k = 8$), and the data is normalized to zero mean and unit variance. The value of k is chosen to be at the same time small enough in order to have realistic computational times, but also high enough to ensure enough inter-cluster disparity. In order to take advantage of the close proximity of artifacts and seeds, Connected Component Analysis (CCA) is performed using a 6-connectivity in 3D for each distinct patch. For each cluster component (i.e. distinct blob for a given cluster), the connectivity with voxels previously labeled in the distance map ($D_{L\neq 0}$) is verified, and unconnected component will be discarded under the label C_0, as shown in Figure 6.

Once unconnected cluster blobs have been rejected, one and only one seed-connected cluster C_i (with $i \in [1; k]$) is selected to be added to the distance map as a new label D_{L+1}, based on a metric $M(C_i)$. According to the literature, the removal of artifacts has a direct effect on the decrease of standard deviation (SD) of the ROI. Following tests with different metrics, we chose to use Eq. 3, representing the change in the SD σ induced by the removal of a given cluster C_i from $P(D_{L=0})$, relative to the cardinality $|C_i|$ of the considered cluster.

$$M(C_i) = \frac{\sigma(P(D_{L=0})) - \sigma(P(\overline{D_{L\neq 0} \cup C_i}))}{|C_i|} \quad (3)$$

The cluster C_i giving the highest value of $M(C_i)$ (and therefore the highest relative change is the SD of $P(D_{L=0})$) is selected, and added in the distance map under the label $L + 1$. Then, new iterations are performed (K-means - CCA - cluster selection - update) until a given criteria. While we tried to use the value of $M(C_i) < 0$ for the selected cluster (representing an increase of the SD of $P(D_{L=0})$) as a criteria, results have shown that such criteria caused and 'early stop', leading

Figure 5: Central axial slice of a given patch. On the left are shown the CT values, in the middle the segmentation of all seeds in the patch, and on the right the segmentation of the primary seed only

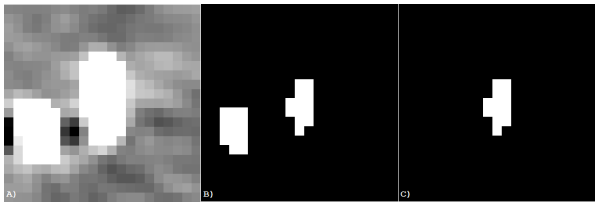


Figure 6: Case of cluster exclusion using connected component analysis. Left: HU value; Center: K-means clustering with k = 8; Right: Retained cluster blobs after 3-D CCA.

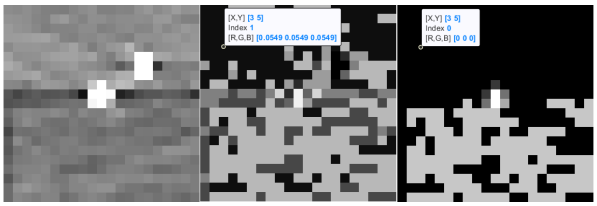
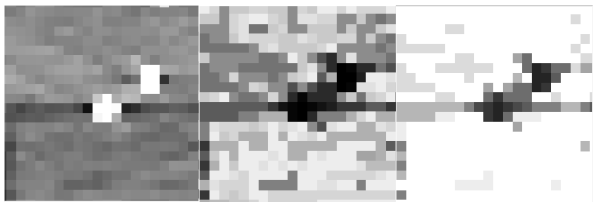


Figure 7: Example of all 21 computed clusters (center) for given CT slice (left). On right are shown retained clusters (X = 12) for correction. The darker the label for a given cluster, the faster the voxel considered was included in the distance map



to the partial correction of artifacts. Therefore, a criteria based on the total number of clustered elements in $P()$ is used. Thus, this iterative process (clustering - CCA - cluster selection) is performed until 80% of P() is clustered. This empirical value of this criteria is chosen to be high to ensure that all possible artifacts are considered.

Once the iterative process is finished, the labels L contained in P are propagated onto the original CT-scan (see Figure 7). Since a specific voxel can appear in one or more distinct patches, it might have different distance label across several patches. This is due to the CCA, where more iterations might have been necessary to reach the considered voxel when starting from one primary seed compared to another. Since the principle idea of the proposed method is to reflect the rapidity at which the considered voxel has been selected regarding the decrease of $\sigma(P(D_{L=0}))$ starting from a given primary seed, keeping only the minimum label during propagation is the most suitable approach. During this merging step, only the labels with a high number of elements are kept (so that combined, they represent 95% of the labeled voxels), and labels with a small number of voxels are merged with bigger labels (according to the mean value of neighbouring labels). From the resulting distance map, only the X first clusters will be selected for correction. This is due to the lack of a stopping criterion for the iterative framework, thus inducing the progressive clustering of background voxels (i.e. artifact-free tissue). For the same reason, no criterion has been found for the selection of the value of X, which has to be determined manually for the time being. Finally, once the clusters to correct are selected, 3-D mean filtering is performed using an evenly-weighted filter of radius 2.

3.3. Dosimetric & statistical analysis

In order to be able to accurately compare the influence on the proposed MAR algorithm on dose calculation, the prostate segmentation performed by the expert on the original CT-scan has been transferred on the corrected CT-scan, hence ensuring that that the exact same volume is used for the calculation of dosimerty parameters. The values of the following parameters have been extracted using VariSeed 9.0 point-source dose calculation before and after correction of the CT-scans:

- D_{90}, the minimal dose received by 90% of the prostate;
- V_{100} and V_{150}, the prostate volumes receiving 100% and 150% of the prescribed dose respectively;
- number of automatically extracted seeds

; For the seed detection, the seed finder tools included in VariSeed was used. As one can explicitly indicate the number of expected seeds to be found, we used the number of seeds detected on the original CT-scan. To interpret the possible changes in values of the parameters D_{90}, V_{100} and V_{150},the following values are computed before and after correction: *i)* mean with the associated standard deviation; *ii)* median value and *iii)* the range of the parameter values. Moreover, the agreement between parameters values before and after correction is assessed by using paired samples Student's t-tests (p significance level: 5%), Spearman correlation coefficient ρ and Bland-Altman analysis (Bland and Altman, 1986).

4. Results

In the previous section, the correction process was described. It has been applied on the dataset, and the same set of parameters was used across all patients, at the exception of the value of X (selection of the X first clusters for correction).

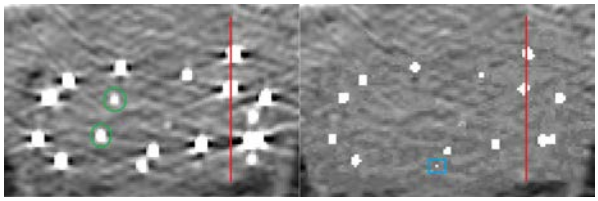
An omnipresent difficulty in the evaluation of MAR algorithms is the lack of both standardized correction validation procedure and the absence of a metric for description and grading of metal artifacts. Four common approaches found in the literature are: *i)* visual assessment of the proposed correction; *ii)* quantification of the noise reduction in a defined ROI, being either a given tissue or region; *iii)* study of the dose calculation differences induced by the MAR algorithm; *iv)* intra- and inter-observer studies on the ease of delineation of tissues, bones or organs. While *ii)*, i.e. the evaluation of the noise reduction in the CT-scan is the most used method for MAR evaluation, in our opinion such evaluation does not fit to this work. This is motivated by the fact that the noise reduction is often quantified by the reduction of the standard deviation of the signal in a given region. The reduction of the standard deviation is used as part of the metric for cluster selection in our proposed MAR algorithm, and therefore one could argue then that such evaluation would be biased. For this work, approaches *i)* and *iii)* were chosen.

Examples of correction are given in Figure 8 where severe artifacts are corrected, and in Figure 9, where minor artifacts are present in the vicinity of the seeds

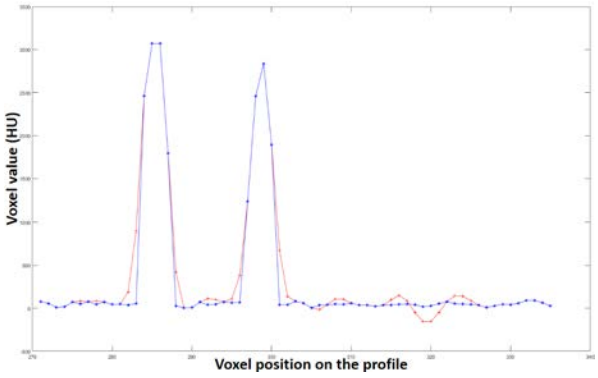
Figure 8: Original and corrected CT slices with highlighted points of interests.

(a) Original CT slice (left) and the associated corrected CT slice (right). Highlighted in green are areas where transaxial Partial Volume Effect (PVE) has been corrected, while the blue highlight indicates a region subject to both transaxial and transverse PVE.

(b) Plot of the HU values associated with the profile indicated by the red line on Figure 8a. In red are represented the original HU values, and in blue the corrected HU values.



(a)



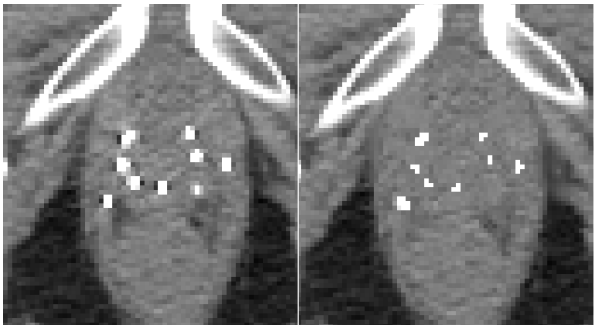
(b)

only. By comparison of the original and corrected CT-scans, it is easy to see that most of the artifacts have been corrected, regardless of their magnitude. While it clearly appears that the CT has been modified in some regions (due to simple interpolation method), one can not easily assess the severity of the corruption in the original image solely by looking at the corrected scan. In Figure 8b are shown the HU values along the profile depicted by the red line. It shows a sharpening of the edges in the close proximity of the seeds, as well as an increased stability in the HU values of prostate tissue.

On the original CT slice (Figure 8a), seeds that were totally removed on the corrected slice are circled in green. The presence of medium to high HU values in the outlined area (200 - 700 HU) is due to the transaxial PVE induced by the presence of the seeds in adjacent slices. This effect inaccurately makes the seed length appear longer that it is in reality, and is therefore corrected.

While the size of the seed highlighted by the blue square might seem to have been wrongly reduced, the HU value of the only voxel kept (in this given slice) was two times higher than the highest HU value adjacent voxels. This is due to the transverse PVE, causing an increase of the HU values in the vicinity of the seeds,

Figure 9: CT slice containing metal artifacts (left) with the associated corrected slice (right). In presence of minor metal artifact, the regions to correct are not overestimated.



leading to an important overestimation of the seed diameter on the CT-scan compared to its physical diameter. It is to note that PVE is hard to identify with commonly used display parameters as they are selected in order to give a good contrast in tissues (thus giving low to non-existent contrast for higher CT numbers). Figure 10 shows the same area with a windowing better suited for increased contrast in high HU values. Using these display parameters, it now appears clearly that the HU values of the highlighted regions are not uniform, in opposition with the impression given by Figure 8a.

After correction, the number of automatically detected seeds ranged from 16 to 97 units (mean 61 ± 22), which is similar to the detection on original data (mean 60 ± 22). However, after correction only two cases shared the exact same detected seed number and location, with respectively 0.12 *mm* and 0.13 *mm* centroid location error, cumulated across all axis. For the other cases, the average absolute seed number difference in seed number is 2.9 ± 2.4 with a median of 1, and with a maximum difference of 8 more detected seeds after correction. In the majority of cases (85%), the number of detected seed was increased after CT-scan MAR process. One can assume that the proposed method, and

Figure 10: CT slice with display parameters selected for a better contrast in medium to high HU values. Highlights correspond to areas of PVE. (Window: 1200; Level: 400)

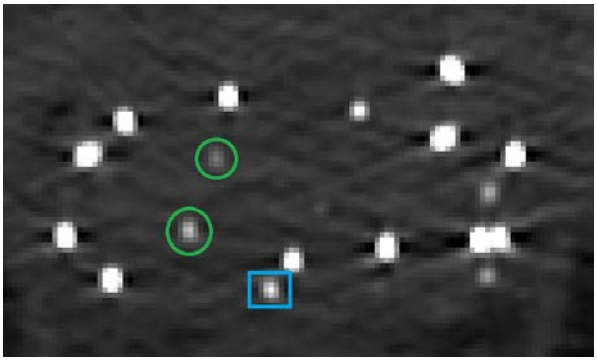


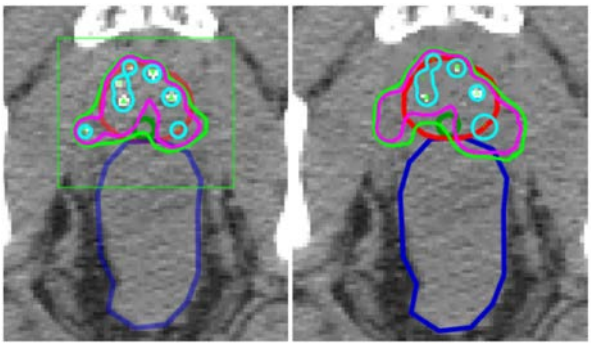
Table 1: Clinical dose calculation parameters differences between original and corrected CT-scans. SD: Standard Deviation; p: paired Student's t-test p-value; ρ: Spearman correlation coefficient; BA: Bland-Altman analysis.

	D_{90} (Gy)		V_{100} (%)		V_{150} (%)		Volume (cc)
	Original	Corrected	Original	Corrected	Original	Corrected	
Mean (SD)	90.43 (35.11)	93.56 (37.05)	80.67 (21.57)	81.71 (21.33)	46.95 (18.45)	50.46 (20.42)	47.51 (19.30)
Median	99.50	100.37	89.73	90.16	44.37	46.55	42.47
Range	13.38 - 163.70	13.92 - 174.19	22.97 - 99.88	23.66 - 100.00	14.62 - 94.747	15.24 - 97.40	10.96 - 89.50
p	0.002		0.020		0.001		
ρ	0.97		0.97		0.91		
BA Bias	-0.93		2.46		0.51		

especially the correction of the PVE, can have a strong influence in seed detection, especially when seeds are close to each other, but its possible benefits are yet to be validated (discussed in Section 5).

Regarding the two cases having the same number of seeds before and after correction, their respective differences in dose calculation parameters are : D_{90} -0.31% / -0.44% , V_{100} 0.08% / -0.28%, V_{150} -0.20% / -0.07%. Then, we can say that results show no significant difference for these two cases. For the 31 other cases, results are shown in Table 1, significant difference is noted across all parameters ($p < 0.05$). In fact, while the Bland-Altman analysis show a bias close to 0, this bias induce small significant difference, especially on the V_{100} dosesimerty parameter. Moreover, the value of the Spearman coefficient ρ being close to one, results show a high correlation between the parameters evaluated before and after MAR. In Figure 11 are shown the isodose lines for a cases where the number of automatically detected of seeds has changed after MAR.

Figure 11: Dosimetry analysis on post-implant brachytherapy CT-scans before(left) and after (right) MAR, where the number of detected seeds was different. The prostate and rectum are outlined in red and blue, respectively. In magenta, green and blue are represented the cumulative isodose line for 240 Gy, 180 Gy and 120 Gy respectively.



5. Discussion

We present in this work a post-processing MAR framework applied on reconstructed CT-scans. The dataset was composed of 33 CT-scans having a high variety regarding the strength of metal artifacts. By visual evaluation, one can say that most of the metal artifacts have been corrected by the proposed framework (more comparisons are shown in Appendix 2). Compared to the uncorrected data, significant difference in dose calculation parameter after the correction of metal artifacts is noted. Although the dose calculation was performed in compliance with the TG-43 point-source formalism, this significant difference is certainly due to the difference in automatically detected seeds. As mentioned previously, the challenge of the quantification and qualification of metal artifacts has yet to be solved. In this section are first discussed limitations of our first approach and then the proposed framework, followed by the challenges faced regarding the evaluation of the dose calculation changes after MAR.

5.1. Autoencoder approach

The poor discrimination between the corrupted-uncorrupted patches extracted in the scope of our first approach was not expected, especially after the accurate reconstruction of patches using an encoder-decoder network. The initial reconstruction errors were present mostly for high HU values (seed centers), and were removed by reducing the HU value resolution for all values above 1500 HU. Regardless, the features extracted from the last encoders' layer were not enabling an accurate classification of the patches by means of SVM. One possible reason can be the extremely high variety regarding the patches' content in both classes. The creation of a new dataset or more classes (such as seed, bone, dark streak, PVE) might have produced better results, given that the spatial resolution is high enough to accurately capture the local data distribution.

In order to learn a deep representation of the data in the latent space, several layers are necessary in the AEs. However, the number of possible cascaded convolutional layers is limited by the input size of the data. As there is only a small distance between seeds, an increase in the input size often induce an even higher variance is the spatial location of high HU values. Also, the higher the input size the lower the ratio of metal artifact voxels over tissue voxels (for the metal artifact class). While the information in adjacent axial slice can help the correction of a given slice, the use of a small input size (5 x 5 x 5 px) induces the consideration of the 2

adjacent axial slices in a given direction (for feature extraction). Knowing that: *1)* the slice thickness of the data is 1.25 *mm*; *2)* the seed length in the its longitudinal direction is 4.8 *mm*; *3)* the transaxial PVE produced by a given seed occurs no further than one slice after the last slice containing the seed; and *4)* in the case of seed whose longitudinal axis is normal to the axial slice, a maximum of about 6 axial slices are sufficient to completely contain all the metal artifacts, one could argue that an increase of the input size above a value of (5 x 5 x 5 *px*) would not add relevant information w.r.t to the correction of metal artifacts originated from the given seed (argument valid outside the scope of the AE training). This motivated the use of a 2.5-D AE, for which an increase of the input size in the saggital and coronal direction were not enforcing an size increase in the axial direction. While the best final classification accuracy was improved (0.73 AUC for 3D, 0.81 AUC for 2.5-D), this approach did not produce results satisfactory enough for further investigation.

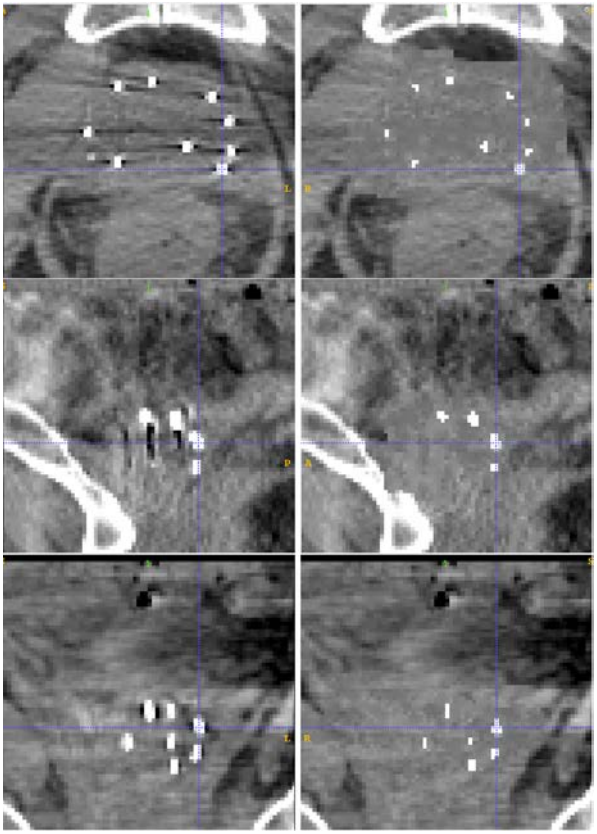
5.2. Clustering approach

The application of clustering methods onto data subject to spatial constraints (with a goal of segmentation) is not a straightforward task. However, due to the absence of ground-truth data and discriminative metric for metal artifact correction, clustering methods were, in our opinion, a suitable approach. In the presented framework, we tried to enforce spatial constraints by iteratively expanding the scope of the data available for clustering by means of connected components analysis, starting from the initial seed segmentation. However, this method has its limitations, which are presented below, associated with the ideas explored to solve them. It is noteworthy that the changes discussed below have not been validated, and therefore have not been used in order to obtain results presented in this work (Girum et al., 2018).

First, a remark concerns the initial seed segmentation by K-mean with $k = 2$. While the resulting segmentation might be not be optimal nor validated, one could provide an accurate segmentation of the seed as the initialization of the distance map. In the proposed method, it is important for the seed area to not be overestimated outside the PVE, and its underestimation does not have negative effects, as long as at least 1 voxel is segmented to represent the seed centroid. In the close future, a brachytherapy seed segmentation method currently under development will be joined with this work in order to combine accurate seed segmentation and MAR.

As visible on the left panel of Figure 12, the clustering might be wrongly extended outside of the prostate, notably in lower density regions and the pelvic bone. This effect is often produced by the propagation of the metal artifact outside of the prostate tissue, hence leading to a positive response of the CCA for this regions. As the HU values of such area are far from the average

Figure 12: Transverse (top), sagittal (middle) and coronal (bottom) CT slices, extracted from a given voxel location both before (left) and after (right) correction.



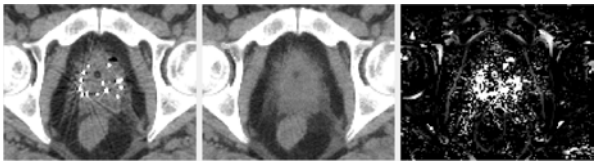
value of prostate tissue, these regions will be selected for correction as soon as they are reached. A simple method to avoid this undesired effect is the integration of the prostate segmentation as a constraint for the correction detection. In the scope of this work, this effect does not have an impact aside from the visual disturbance, and is left uncorrected for clarity purposes regarding the algorithm limitations.

A second label refining option is the final sorting of the labels w.r.t Eq.3. It was noted that this step, while inducing a loss of the initial semantic role of the distance map, generally reordered few labels, being mostly minor dark band artifacts in seed vicinity. However, this step often produces a worse correction in presence of the aforementioned effect (out of prostate correction). Indeed, while having a relatively high label in the distance map, regions outside of the prostate will inherently rank in the first positions according to a sorting w.r.t Eq. 3.

An explored option was the definition of a model to map each voxel to a real number in the domain $[0; 1]$, representing the degree of corruption of a voxel. The main advantage of this approach was the mapping of the data to values ranging from 0 to 1. In fact, such value could then be used directly as a local weighting factor for the interpolation of corrupted data. Thus, using such

representation of the corruption, one could directly replace the currently used evenly-weighted 3-D mean filter of arbitrary size centered at a given location with a filter whose weights are those of the CCS at the same location. An example of result for this method is given in Figure 13. For this approach, the corruption detection method was similar to the one rpesented in (Basran et al., 2011). We tried to transpose the generation of this idea to the presented framework by also taking into account the distance map and clusters' mean without success. Such scoring could remove the need for the manual selection of X in the current framework (X being the last label considered for correction).

Figure 13: Original (left), Corrected (middle) and CCS map (right) slices for a particular case. The seeds have also been removed during the process.



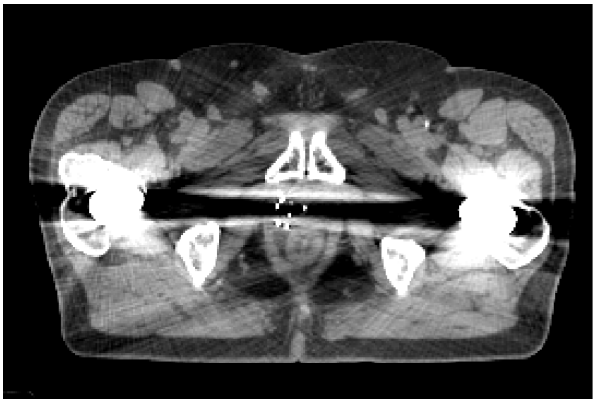
Finally, a remark is given on the presence of metal artifacts due to metallic objects exterior to the prostate. As visible on the top-left image from Figure 12, the presence of such artifacts does not only strengthen the artifacts originating form BT seeds, it also makes both the metal artifact detection and data interpolation steps more difficult. The presence of a such a wide dark streak will at the same time induce a bias on the estimation of the standard deviation in HU values inside the prostate (and therefore on the estimation of clusters to correct) and also introduce a similar bias in the interpolation of missing data. The obtained correction, while acceptable in cases displayed in Figure 12, does not have any noticeable impact in cases such as figure 14. In this case, the use of a dedicated MAR for large metallic implants (as reviewed in Section 2.3) before BT seeds MAR correction is mandatory.

5.3. Dosimetry evaluation & Perspectives

In Section 4, consistent dose calculation before and after MAR have been shown. In order to take this work further, the authors would like to follow a different approach for the evaluation of the impact of the proposed MAR framework on dose calculation parameters. As mentioned in Section 2.2, no TPS using MBDCA can currently perform the dose calculations for LDR-BT seeds, such as the I-125 seed used to treat all patient in the presented dataset. Therefore, there is currently no way to evaluate the dose calculation differences due to, and only to, the MAR algorithm.

However, it has been shown in this work that the use of our MAR framework induced changes in the number of seeds automatically detected by VariSeed seed

Figure 14: CT slice of a case for which the proposed framework had no noticeable impact



finder tool in a significant number of cases. In this work, the cases forming the dataset are from clinical cases of primary permanent prostate implant and from salvage permanent prostate implant. This leads to a high variability regarding the true number of implemented seed. While the number and location of the seed are determined before the implant, this values can, and most probably will, change due to phenomena linked to the time elapsed between the implantation of the seeds and the imaging of the patients' prostate after the implant. In fact, changes in seed location can be noted over time. Moreover, such displacements sometimes lead to the rejection of the seed through the urethra. Therefore, after a given time, both seed number and location will have changed. If one want to perform a study on the changes in seed detection after MAR, common method to acquire ground truth data for seed number and location are the use of phantoms and experts annotation on CT-scan. While the first have the downside of not being a clinical case, the second can often be performed for few cases only, as the task extensive time for one expert. However, we would like to take this work further by following one of the two mentioned option.

Finally, we will continue our work in order to improve the proposed framework, with the goal of reaching a fully automated process.

6. Conclusions

While MAR methods have been a strong point of interest in the past year, a small number of studies present methods for the correction of metal artifacts due to the presence of brachytherapy sources, and to the best of our knowledge, only one approach is not based on raw sinogram data. In this work has been presented a framework for MAR on reconstructed CT-scans. The proposed solution is based on the strong spatial relationship between metal artifacts and BT seeds. It has been shown that the proposed MAR algorithm did not induce

a significant difference on dose parameters when the automatic seed detection was consistent, while small significant changes were noted when the number of detected seeds changed. We hope to extend this work to a larger framework, where post-implant CT-scans could be processed for seed detection, segmentation, and MAR. In our opinion, such framework could increase dose calculation accuracy, and ease the challenge of automatic prostate segmentation, especially in the presence of metal artifacts.

7. Acknowledgements

First and foremost, I would like to express my sincerest gratitude towards my supervisors, Dr. Alain Lalande and Pr. Gilles Crhange, for their guidance, patience, support and kindness throughout the entirety of my Master Thesis, regardless of the challenges faced.

Secondly, my deepest appreciation goes to PhD student Kibrom Berihu Girum for taking the time to share with me his knowledge and answer all my questions, always with a smile.

I would also like to tanks all my friends from MAIA for the great time we had, with a special mention for Lavsen Dahal who helped me review this work.

Finally, *un grand merci* to my family for their unconditional love and support despite all the hiccups induced by the debatable decisions I took over the years. *Je vous aime*.

References

Ahnesjö, A., 1989. Collapsed cone convolution of radiant energy for photon dose calculation in heterogeneous media. Medical physics 16, 577–592. doi:10.1118/1.596360.

Andersson, K.M., Nowik, P., Persliden, J., Thunberg, P., Norrman, E., 2015. Metal artefact reduction in ct imaging of hip prostheses: an evaluation of commercial techniques provided by four vendors. The British journal of radiology 88, 20140473. doi:10.1259/bjr.20140473.

Axente, M., Paidi, A., Von Eyben, R., Zeng, C., Bani-Hashemi, A., Krauss, A., Hristov, D., 2015. Clinical evaluation of the iterative metal artifact reduction algorithm for ct simulation in radiotherapy. Medical physics 42, 1170–1183. doi:10.1118/1.4906245.

Baldi, P., 2012. Autoencoders, unsupervised learning, and deep architectures, in: Proceedings of ICML workshop on unsupervised and transfer learning, pp. 37–49. doi:10.1.1.296.3397.

Barrett, J.F., Keat, N., 2004. Artifacts in ct: recognition and avoidance. Radiographics 24, 1679–1691. doi:10.1148/rg.246045065.

Basran, P.S., Robertson, A., Wells, D., 2011. Ct image artifacts from brachytherapy seed implants: A postprocessing 3d adaptive median filter. Medical physics 38, 712–718. doi:10.1118/1.3539648.

Bazalova, M., Beaulieu, L., Palefsky, S., Verhaegen, F., 2007. Correction of ct artifacts and its influence on monte carlo dose calculations. Medical physics 34, 2119–2132. doi:10.1118/1.2736777.

Beaulieu, L., Tedgren, Å.C., Carrier, J.F., Davis, S.D., Mourtada, F., Rivard, M.J., Thomson, R.M., Verhaegen, F., Wareing, T.A., Williamson, J.F., 2012. Report of the task group 186 on model-based dose calculation methods in brachytherapy beyond the tg-43 formalism: current status and recommendations for clinical implementation. Medical physics 39, 6208–6236. doi:10.1118/1.4747264.

Bland, J.M., Altman, D., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. The lancet 327, 307–310. doi:10.1016/S0140-6736(86)90837-8.

Boas, F.E., Fleischmann, D., 2012. Ct artifacts: causes and reduction techniques. Imaging Med 4, 229–240. doi:10.2217/iim.12.13.

Bolstad, K., Silje, F., Aadnevik, D., Dalehaug, I., Vetti, N., 2018. Metal artifact reduction in ct, a phantom study: subjective and objective evaluation of four commercial metal artifact reduction algorithms when used on three different orthopedic metal implants. Acta Radiologica doi:10.1177/0284185117751278.

Briesmeister, J.F., et al., 2000. Mcnptm-a general monte carlo n-particle transport code. Version 4C, LA-13709-M, Los Alamos National Laboratory , 2.

Chang, Y.B., Xu, D., Zamyatin, A.A., 2012. Metal artifact reduction algorithm for single energy and dual energy ct scans, in: Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2012 IEEE, IEEE. pp. 3426–3429. doi:10.1109/NSSMIC.2012.6551781.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine learning 20, 273–297. doi:10.1007/BF00994018.

Créhange, G., Hsu, I.C., Chang, A.J., Roach, M., 2017. Salvage prostate brachytherapy for postradiation local failure, in: Management of Prostate Cancer. Springer, pp. 287–302.

Giantsoudi, D., De Man, B., Verburg, J., Trofimov, A., Jin, Y., Wang, G., Gjesteby, L., Paganetti, H., 2017. Metal artifacts in computed tomography for radiation therapy planning: dosimetric effects and impact of metal artifact reduction. Physics in Medicine & Biology 62, R49. doi:10.1088/1361-6560/aa5293.

Girum, K.B., Lalande, A., Quivrin, M., Bessières, I., Pierrat, N., Martin, E., Cormier, L., Petitfils, A., Cosset, J.M., Créhange, G., 2018. Inferring postimplant dose distribution of salvage permanent prostate implant (ppi) after primary ppi on ct images. Brachytherapy 17, 866–873. doi:10.1016/j.brachy.2018.07.017.

Gjesteby, L., De Man, B., Jin, Y., Paganetti, H., Verburg, J., Giantsoudi, D., Wang, G., 2016. Metal artifact reduction in ct: where are we after four decades? IEEE Access 4, 5826–5849. doi:10.1109/ACCESS.2016.2608621.

Healthcare, P., 2012. Metal artifact reduction for orthopedic implants (o-mar). White Paper, Philips CT Clinical Science, Andover, Massachusetts doi:10.2214/AJR.16.17684.

Huang, J.Y., Kerns, J.R., Nute, J.L., Liu, X., Balter, P.A., Stingo, F.C., Followill, D.S., Mirkovic, D., Howell, R.M., Kry, S.F., 2015. An evaluation of three commercially available metal artifact reduction methods for ct imaging. Physics in Medicine & Biology 60, 1047. doi:10.1088/0031-9155/60/3/1047.

Karaiskos, P., Papagiannis, P., Sakelliou, L., Anagnostopoulos, G., Baltas, D., 2001. Monte carlo dosimetry of the selectseed interstitial brachytherapy seed. Medical physics 28, 1753–1760. doi:10.1118/1.1384460.

Karimi, S., Cosman, P., Wald, C., Martz, H., 2012. Segmentation of artifacts and anatomy in ct metal artifact reduction. Medical physics 39, 5857–5868. doi:10.1118/1.4749931.

Kawrakow, I., 2001. The egsnrc code system, monte carlo simulation of electron and photon transport. NRCC Report Pirs-701 .

Lee, Y.H., Park, K.K., Song, H.T., Kim, S., Suh, J.S., 2012. Metal artefact reduction in gemstone spectral imaging dual-energy ct with and without metal artefact reduction software. European radiology 22, 1331–1340. doi:10.1007/s00330-011-2370-5.

Lloyd, S., 1982. Least squares quantization in pcm. IEEE transactions on information theory 28, 129–137. doi:10.1109/TIT.1982.1056489.

Meyer, E., Raupach, R., Lell, M., Schmidt, B., Kachelrieß, M., 2010. Normalized metal artifact reduction (nmar) in computed tomography. Medical physics 37, 5482–5493. doi:10.1118/1.3484090.

Meyer, E., Raupach, R., Lell, M., Schmidt, B., Kachelrieß, M., 2012. Frequency split metal artifact reduction (fsmar) in computed tomography. Medical physics 39, 1904–1916. doi:10.1118/1.3691902.

Mouton, A., Megherbi, N., Van Slambrouck, K., Nuyts, J., Breckon, T.P., 2013. An experimental survey of metal artefact reduction in computed tomography. Journal of X-ray Science and Technology 21, 193–226. doi:10.3233/XST-130372.

Nag, S., Erickson, B., Thomadsen, B., Orton, C., Demanes, J.D., Petereit, D., Society, A.B., 2000. The american brachytherapy society recommendations for high-dose-rate brachytherapy for carcinoma of the cervix. International Journal of Radiation Oncology* Biology* Physics 48, 201–211.

Nath, R., 1995. Recommendations of the aapm radiation therapy committee task group no. 43. Med. Phys. 22, 209–234. doi:10.1118/1.597458.

Pessis, E., Campagna, R., Sverzut, J.M., Bach, F., Rodallec, M., Guerini, H., Feydy, A., Drapé, J.L., 2013. Virtual monochromatic spectral imaging with fast kilovoltage switching: reduction of metal artifacts at ct. Radiographics 33, 573–583. doi:10.1148/rg.332125124.

Rivard, M.J., Ballester, F., Butler, W.M., DeWerd, L.A., Ibbott, G.S., Meigooni, A.S., Melhus, C.S., Mitch, M.G., Nath, R., Papagiannis, P., 2017. Supplement 2 for the 2004 update of the aapm task group no. 43 report: joint recommendations by the aapm and gec-estro. Medical physics 44, e297–e338. doi:10.1002/mp.12430.

Rivard, M.J., Butler, W.M., DeWerd, L.A., Huq, M.S., Ibbott, G.S., Meigooni, A.S., Melhus, C.S., Mitch, M.G., Nath, R., Williamson, J.F., 2007. Supplement to the 2004 update of the aapm task group no. 43 report. Medical physics 34, 2187–2205. doi:10.1118/1.2736790.

Rivard, M.J., Coursey, B.M., DeWerd, L.A., Hanson, W.F., Saiful Huq, M., Ibbott, G.S., Mitch, M.G., Nath, R., Williamson, J.F., 2004. Update of aapm task group no. 43 report: A revised aapm protocol for brachytherapy dose calculations. Medical physics 31, 633–674. doi:10.1118/1.1646040.

Rivard, M.J., Venselaar, J.L., Beaulieu, L., 2009. The evolution of brachytherapy treatment planning. Medical physics 36, 2136–2153. doi:10.1118/1.3125136.

Rozet, F., Hennequin, C., Beauval, J.B., Beuzeboc, P., Cormier, L., Fromont-Hankard, G., Mongiat-Artus, P., Ploussard, G., Mathieu, R., Brureau, L., et al., 2018. Recommandations françaises du comité de cancérologie de l'afu actualisation 2018 2020: cancer de la prostate. Progrès en urologie 28, S79–S130. doi:10.1016/j.purol.2018.08.011.

Shiraishi, Y., Yamada, Y., Tanaka, T., Eriguchi, T., Nishimura, S., Yoshida, K., Hanada, T., Ohashi, T., Shigematsu, N., Jinzaki, M., 2016. Single-energy metal artifact reduction in postimplant computed tomography for i-125 prostate brachytherapy: Impact on seed identification. Brachytherapy 15, 768–773. doi:10.1016/j.brachy.2016.07.006.

Siegel, R.L., Miller, K.D., Jemal, A., 2019. Cancer statistics, 2019. CA: a cancer journal for clinicians doi:10.3322/caac.21442.

Steger, C., 1998. An unbiased detector of curvilinear structures. IEEE Transactions on pattern analysis and machine intelligence 20, 113–125. doi:10.1109/34.659930.

Takahashi, Y., Mori, S., Kozuka, T., Gomi, K., Nose, T., Tahara, T., Oguchi, M., Yamashita, T., 2006. Preliminary study of correction of original metal artifacts due to i-125 seeds in postimplant dosimetry for prostate permanent implant brachytherapy. Radiation medicine 24, 133–138. doi:10.1007/BF02493280.

Veldkamp, W.J., Joemai, R.M., van der Molen, A.J., Geleijns, J., 2010. Development and validation of segmentation and interpolation techniques in sinograms for metal artifact suppression in ct. Medical physics 37, 620–628. doi:10.1118/1.3276777.

Williamson, J.F., 1987. Monte carlo evaluation of kerma at a point for photon transport problems. Medical physics 14, 567–576.

Xu, C., Verhaegen, F., Laurendeau, D., Enger, S.A., Beaulieu, L., 2011. An algorithm for efficient metal artifact reductions in permanent seed implants. Medical physics 38, 47–56. doi:10.1118/1.3519988.

Yang, Q., Peng, S., Wu, J., Ban, X., He, M., Xie, C., Zhang, R., 2015. Spectral ct with monochromatic imaging and metal artifacts reduction software for artifacts reduction of 125 i radioactive seeds in liver brachytherapy. Japanese journal of radiology 33, 694–705.

Zhou, C., Inanc, F., 2002. Integral-transport-based deterministic brachytherapy dose calculations. Physics in Medicine & Biology 48, 73. doi:10.1088/0031-9155/48/1/306.

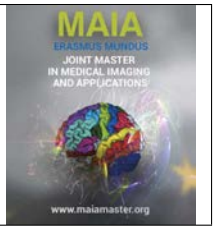
Appendix 1: TG-43 definitions

Following are the definition of terms used in the TG-43 formalism:

- A source is defined as any encapsulated radioactive material that may be used for brachytherapy. There are no restrictions on the size or on its symmetry.
- A point source is a dosimetric approximation whereby radioactivity is assumed to subtend a dimensionless point with a dose distribution assumed to be spherically symmetric at a given radial distance r. The influence of inverse square law, for the purpose of interpolating between tabulated transverse-plane dose-rate values, can be calculated using 1/r2.
- The transverse-plane of a cylindrically symmetric source is that plane which is perpendicular to the longitudinal axis of the source and bisects the radioactivity distribution.
- A line source is a dosimetric approximation whereby radioactivity is assumed to be uniformly distributed along a 1D line-segment with active length L. While not accurately characterizing the radioactivity distribution within an actual source, this approximation is useful in characterizing the influence of inverse square law on a sources' dose distribution for the purposes of interpolating between or extrapolating beyond tabulated TG-43 parameter values within clinical brachytherapy treatment planning systems.
- A seed is defined as a cylindrical brachytherapy source with active length, L, or effective length, Leff less than or equal to 0.5 cm.

Medical Imaging and Applications

Master Thesis, June 2019



Transfer Learning in Medical Imaging

Kenechukwu Henry Ngige, Mario Molinara

Faculty of Engineering, University of Cassino and Southern Lazio, Cassino Italy

Abstract

Convolutional neural networks (CNNs) have become one of the state of the art methods in image classification in various applications. Training a convolutional neural network from scratch is sometimes difficult because it requires a large amount of training data to ensure adequate convergence. In medical image classification where the training data might be small in some situation, transfer learning using CNNs is often applied. To get a large set of medical images for knowledge transfer is sometimes difficult. Alternative such as a large set of labeled natural images used on the state of the art CNNs are generally used. One of the technique of transfer learning extracts generic image features from the natural images and these can be applied in feature extraction in smaller dataset. Another technique of transfer learning called fine tuning is to train a CNN from a set of weights pre-trained on large dataset. However, the difference between natural and medical images may not be suitable in some knowledge transfer. In this thesis work, the two transfer learning mode which is based on feature extraction and fine tuning are compared with training from scratch using five selected state of the art pre-trained CNN models which include VGG19, ResNet50, InceptionResNetV2, InceptionV3 and NASNetLarge. The fully connected layers on top of the base models were removed and a new classifier on top of the base models was designed. Performance was evaluated on two publicly available datasets, INbreast dataset for breast mass classification and HAM10000 dataset for skin lesion classification. The results showed that transfer learning mode of fine tuning performed better than the other two training approaches in most pre-trained CNN models in breast mass classification, training from scratch performed better than the other two training approaches in most pre-trained CNN models in skin lesion classification, transfer learning mode of feature extraction gave the least performance in all pre-trained CNN models in both classification and in overall, transfer learning mode of fine tuning produced the best performance in both classification.

Keywords: Convolutional Neural Networks, Transfer Learning, Training from Scratch, Breast Mass, Skin Lesion

1. Introduction

1.1. Breast Cancer

Breast cancer is one of the most frequent cancer cases affecting women in the world with an estimated 1.67 million new cancer cases diagnosed in 2012 and the second cause of cancer deaths in women (Ferlay et al., 2013). According to World Health Organization (WHO), it is the fifth most common cause of death (Clemmesen, 1948). In Japan, breast cancer is the most frequent cancer among women in recent years, and the number of patient increases year by year (Matsuda et al., 2014). In Jordan, breast cancer constitutes around one third of all malignancies among females and about 15% of these are due to genetic origin (Abdel-Razeq et al., 2018). Among all races, Chinese women are more prone to develop breast cancer followed by Indian and Malaysian women (Ting et al., 2019). Breast cancer occurs due to damage of genes that regulate the growth and differentiation of cells and this makes them to grow and multiply in an uncontrolled manner. Symptoms of breast cancer include: breast lump, changes to the skin over the breast, changes in breast size or shape, and nipple abnormalities. Studies have shown that there are high chances of survival if it is diagnosed at an early stage (Howlader et al., 2013).

Different factors can increase women risk of breast cancer. Some of the factors that increase the risk of breast cancer include:

- Age: The risk of breast cancer tends to increase with age. This implies that older women are prone to have breast cancer than younger women.
- Previous Ionizing Radiation: Studies have shown that radiotherapy in long term survivors of Hodgkin's disease are prone to develop breast cancer. Young women who received chest radiation with high doses showed an increased risk of breast cancer compared to patients with lower doses (Travis et al., 2005).
- Hormonal Factors: Breast cancer is a hormone related disease. Estrogen has shown to induce and promote mammary gland tumors. Factors that change sex hormone levels tends to increase the risk of breast cancer.
- Hormone Replacement Therapy: Women who suffer from menopausal symptoms at early age caused by reduced estrogen and progesterone levels usually receive hormone replacement therapy. Hormone replacement therapy tends to increase the risk of breast cancer especially among women using combined estrogen and progesterone (Beral et al., 2007).
- Breast density: Dense breast on mammography is associated with an increased risk of breast cancer (Boyd et al., 1995), (Boyd et al., 2002).
- Genetic Factors: Women in which there is a history of breast cancer in their close families are prone to develop breast cancer. There is increased risk of breast cancer if many relatives were affected or if relatives were diagnosed at younger age. High-penetrant germline mutations in BRCA1 (Easton et al., 1995) and BRCA2 (Wooster et al., 1995) which are the two tumor-suppresor genes accounts for about 20% to 25% of inherited breast cancer cases and tends to increase the risk of developing breast cancer at 40% to 80% (Ghoussaini et al., 2013).

The two important markers for breast cancer are breast mass and micro-calcification. Breast masses are more difficult to detect than micro-calcifications due to their blurred features and poor contrast. Breast masses are visually characterized by gray to white regions in the breast area of mammograms, and their shapes are mainly described as oval, irregular or lobulated with boundaries that can be circumscribed, obscured, ill-defined or spiculated (Oliver et al., 2010), (Tang et al., 2009). Breast mass is either benign or malignant. A breast mass with round or oval shape, circumscribed margin and low density has a high probability of being benign, while a breast mass with irregular shape, spiculated margin and high-density has a high probability of being malignant (Grimm et al., 2014). Large variability in describing masses is still reported (Boyer et al., 2013) and thus automatic classification of breast mass is a potential benefit in supporting radiologist final decision. Benign mass is not cancerous because it forms a pseudocapsule that prevents the tumor growth from invading the surrounding normal tissues. In contrast, a malignant breast mass is cancerous because it does not have a pseudocapsule and they tend to invade the surrounding tissues.

Mammography is the main imaging technology used to diagnose breast cancer. The x-ray radiation in mammography can show breast cancer because of the different x-ray absorption rates between the normal and abnormal areas of the breast. The tumors in mammogram images can appear as masses or micro-calcifications. The dense tissue in mammogram images can also look like a mass and may sometimes cover the mass. This makes mammography less sensitive in some situations. In 2011, breast tomosynthesis as a new mammography breast imaging technology was approved by US Food and Drug Administration (FDA). In tomosynthesis, multiple x-ray images are taken from different angles and reconstruction to a video. This makes tomosynthesis to be more helpful to radiologists in identifying various abnormalities in the breast because it avoids the overlay of dense tissue and mass. Also contrasted-enhanced (CE) digital mammography provides more diagnostic accuracy than mammography in dense breasts but it is not easily affordable due to high cost and high levels of radiation (Lewis et al., 2017). Mammography is the widely used imaging technology to diagnose breast cancer because it is more affordable than tomosynthesis. Mammograms are procured in two standard orientations: Craniocaudal (CC) and Medio-lateral-oblique (MLO) views during screening. Figure 1 is an example of the CC and MLO views of the mammogram of two breasts from the same patient.

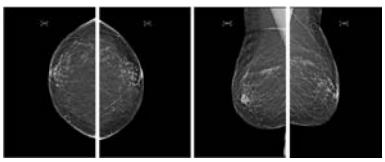


Figure 1: Craniocaudal and Medio-lateral-oblique view of Mammogram

1.2. Skin Cancer

A skin lesion is an abnormal lump, bump, or sore on the skin. A skin lesion on a human body is shown in Figure 2.

Skin lesions include melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, vascular lesion etc. Some skin lesions are benign while some such as melanoma, basal cell carcinoma and squamous cell carcinoma are cancerous. The world health organization (WHO) reported a rapid increase of skin cancer cases (INTERSUN, 2003). Between 2 to 3 million cases of non-melanoma cancer

Figure 2: Skin lesion on a human body

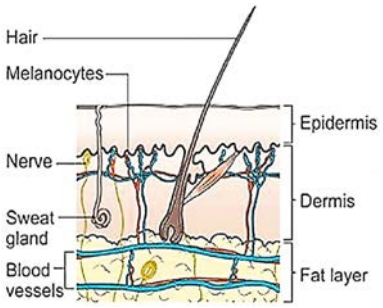


Figure 3: A normal skin anatomy

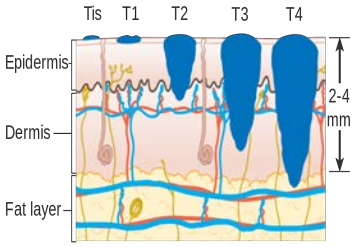


Figure 4: Stages of melanoma

and 132,000 melanoma cancers are reported annually worldwide (Baldwin and Dunn, 2013).

Melanoma is the most dangerous type of skin cancer and the cause of 75% of skin cancer related deaths in the world. It is generally known to occur in the skin but occurrence in the eyes, nasal passages is also possible (Satheesha et al., 2017). Melanoma affects the melanocyte cells which are responsible for producing the skin pigment called melanin. Skin with lack of melanin have more chances of sunburn because of the ultra-violet rays from the sun. The excess sunburn then leads to melanoma. A skin affected by melanoma can be identified by careful observation of the skin area by a dermatologist (a doctor specialized in skin diseases). Classical clinical algorithms such as ABCD - Asymmetry, Border, Color and Diameter (Stolz, 1994), ABCDE - Asymmetry, Border, Color, Diameter and Evolution (Blum et al., 2003), Menzies method (Menzies et al., 1996) and the seven-point checklist (Argenziano et al., 1998) are normally used for the diagnosis of melanoma skin lesion. A normal skin anatomy and the stages of melanoma as defined by Cancer Research UK is shown in Figure 3 and Figure 4 respectively. The stages of melanoma is mainly grouped into five stages.

- Stage 0 (Tis): It can be referred to as the initial stage of melanoma generally called insitumelanoma. It is the first stage of melanoma. Abnormal melanocytes occur in the top layer of the skin. Melanoma in this stage is 100% curable.
- Stage 1 (T1): The melanoma in this stage have spread into the skin but only in the epidermis layer of the skin. The depth is less than 1mm. The patient can be cured through surgical procedure at this stage.
- Stage 2 (T2): The melanoma lesion is between 1mm and 2mm in depth. The patient can still be cured through surgical procedure at this stage.
- Stage 3 (T3): The lesion is between 2mm and 4mm in depth in this stage. The cancer have spread to the lymph nodes but still localized. The patient can still be cured through advanced surgery and post-surgical care but the survival rate is less.
- Stage 4 (T4): The lesion is more than 4mm in depth. The cancer have spread from its primary site to other organs and lymph nodes. The survival rate is very low among patients.

From the above stages, it can be seen that detection and diagnosis of melanoma at the early stage helps in achieving efficient and effective treatment.

Dermoscopy technique which can also be referred to as surface skin microscopy or dermatoscopy is a noninvasive method that allows the in vivo evaluation of micro structures of the epidermis, dermoepidermal junction and the papillary dermis of the skin with the aid of a dermoscope. It is generally performed by a dermatologist. Cascinelli et al. (1987) performed the first pilot study with computer-aided dermoscopic diagnosis using digital slides. Dermoscopy is performed by the application of a gel on the skin and a dermatoscope is then used to obtain a magnified image. The magnified skin images provide color, pattern and structure. This enables the dermatologists to identify the type of skin lesion and help in the diagnosis (Mayer et al., 1997). Other techniques that are used to acquire skin lesion images are high-frequency ultrasound (Vogt and Ermert, 2007), nevoscopy (Dhawan et al., 1984), acoustic microscopy (Tittmann et al., 2013), trans-illumination light microscopy (D'Alessandro and Dhawan, 2012), and 3D high-frequency skin ultrasound images (Pereyra et al.). These techniques construct 3D volumes and estimate the depth of skin lesions for accurate diagnosis. These techniques are not easily available and are more expensive. This makes dermoscopy to be the most widely and affordable technique in the diagnosis of skin lesion.

1.3. Deep Learning

Machine learning have witnessed various developments that lead to a lot of interest from industry, academic and medical domain. This is as a result of breakthrough in artificial neural network generally called deep learning. Deep learning consists of a set of techniques that enable computers to identify complicated patterns in a dataset and their models form the state-of-art used in a wide variety of problems in computer vision. A neural network consist of connected computational units called neurons which are arranged in layers. Data enters the network through the input layer followed by one or more hidden layers which transforms the data as it flows through and ends at an output layer that produces the neural network's predictions. During training, the strength of the neurons are tuned until the patterns identified by the network result in good predictions for the training data. The network use the patterns that are learned to make predictions on new or unseen data.

The basic form of a neural network which is called the feedforward neural networks are parametrized mathematical functions that maps an input to an output and then fed through a number of nonlinear transformations. A neural network is trained by changing its weights to optimize the outputs of the network. This is done by using an optimization algorithm called gradient descent on a loss function which measures the correctness of the outputs. As training data is fed through the network, the gradient of the loss function is computed with respect to every weight using the chain rule and the loss is reduced by changing these weights using the gradient descent. In deep learning, computer learns features and representations directly from the input data. The main characteristic of deep learning is their ability of feature learning whereby representations of data are automatically learned. This is the main difference between deep learning and other classical machine learning.

Deep learning started to be widely known in computer vision when neural networks began outperforming other methods on high-profile image analysis. The most outstanding is the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) when a deep learning model, a convolutional neural network have the second best error rate on the image classification task (Krizhevsky et al., 2012). Before then, to enable computers to recognize objects in natural images was considered to be a difficult task but now convolutional neural networks have exceeded even the human performance on the ILSVRC and arrived the level where ILSVRC classification task is absolutely solved. The deep learning techniques have become a standard for various computer vision tasks. They are not limited to image analysis but also to other areas like natural language processing (Peters et al., 2018), speech recognition and synthesis (Xiong et al., 2018). The wide scope of deep learning has led to improvements in the entire field of machine learning because in some tasks, classical machine learning like support vector machine are now incorporated with deep learning to improve their performance. This has made it to be one of the interesting areas of study world-wide and also offers lucrative job opportunities. People with competence in deep learning are now sought-after in industry, academic and medical domain.

In deep learning, various new methods are developed that will be able to solve problems by learning from experiences. The goal is to create models that can generalize well and deliver accurate predictions on the new or unseen data. The generalization ability of the model is estimated during training using a separate data called the validation data and used as feedback for additional tuning of the model. After several iterations of training and tuning, the final model is evaluated on a test data which is used to show how the model will perform when given a new or unseen data.

In healthcare, large amounts of data which contains valuable information and signals are increasing in a rapid rate that traditional methods of analysis finds it difficult to analyze and process. Deep learning have offered tremendous aid in such amounts of data. In healthcare practice, deep learning is not only used in medical image analysis but also in one-dimensional bio-signal analysis (Ganapathy et al., 2018), analysis of electronic health records (Shickel et al., 2017), stratified care delivery (Vranas et al., 2017), prediction of medical events e.g. cardiac arrests (Kwon et al., 2018), survival analysis (Katzman et al., 2018), aid in therapy selection and pharmacogenomics (Kalinin et al., 2018).

In image recognition, convolutional neural network (CNN) which is a kind of neural network is generally adopted for image classification because of its powerful way to learn useful representations of images and other structured data. The CNN consists of multiple layers of neural connections with minimal systematic processing. The input to a CNN is organized in a grid structure and connected through layers that preserve their relationships with each layer connected to the previous and next layer. CNN's architecture is mainly composed of convolutional, pooling and fully connected layer as shown in Figure 5.

The main function of a convolutional layer is to detect edges, lines and local patterns. Filter operators called convolutions are adopted in the convolution layer and it represents the multiplication of local neighbours from a specified pixel by a certain array of kernel. The extraction of features such as edges are through the kernels. The main function of the pooling layer is to give some translational invariance to the CNN and the fully connected layer act as a classifier to the CNN. The multilayer architecture and local connection in CNN can extract multi-level local features in image data. CNNs enables learning highly representative and hierarchical image features from a large set of training images. The

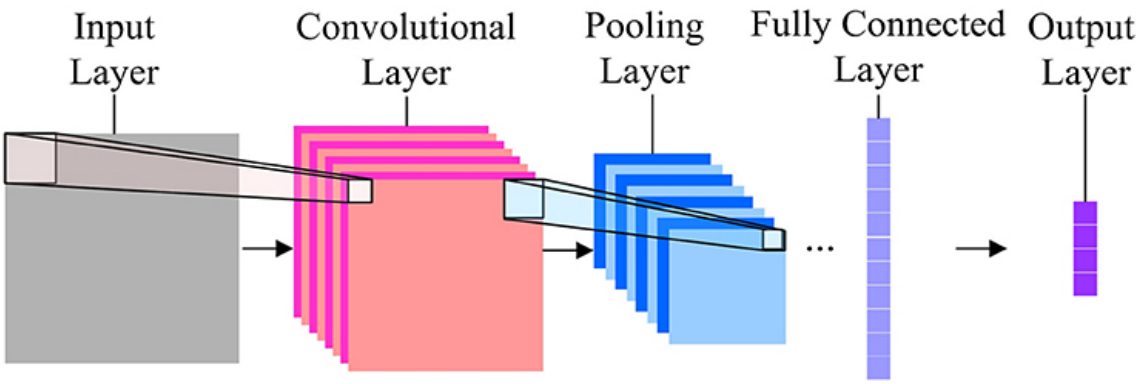


Figure 5: A typical CNN architecture

main power of CNNs exists in its deep architectures (Simonyan and Zisserman, 2014), (Szegedy et al., 2015) which makes CNNs to automatically learn mid-level and high-level abstractions from raw image data. In image recognition, deep CNNs have achieved a great success due to the availability of large annotated data (Deng et al., 2009) and fast graphics processing units - GPUs (Raina et al., 2009). In the field of medical imaging, acquiring data that is annotated as natural images from ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is very difficult. ImageNet (Deng et al., 2009) offers a large database of more than 1.2 million categorized natural images of 1000 classes. CNN models trained on this database help in significantly improving many image classification problems using other datasets. When such large amount of data are not available, using a very limited number of medical data to train the deep CNNs usually causes overfitting and convergence problems.

An approach to exploit deep CNNs on small dataset is transfer learning. Transfer learning is a technique in which the information obtained by a trained model is re-used on another task. In transfer learning, the deep CNNs is first trained on a large image dataset like ImageNet. Natural images and medical images are different but conventional image descriptors developed for object recognition in natural images such as the scale-invariant feature transform - SIFT (Lowe, 2004) and the histogram of oriented gradients - HOG (Dalal and Triggs, 2005) have been widely used for detection in medical image analysis. The two main strategies of which transfer learning can be applied in image classification are based on feature extraction and fine tuning.

In transfer learning mode of feature extraction, the pre-trained CNN models are used as feature extractors. Some previous studies showed that generic descriptors extracted from pre-trained CNN models are effective in recognizing and detecting objects from natural images (Oquab et al., 2014). CNN models that are trained using natural image dataset or which can also be medical dataset can be applied to a new medical task. The pre-trained CNN models are used as feature generators to extract features from the input images and these extracted features are used to train a new classifier such as neural network classifier and support vector machine classifier. Bar et al. (2015) used a pre-trained CNN model as feature extractor in chest pathology detection.

In transfer learning mode of fine tuning, a CNN is trained from a set of weights pre-trained using other large dataset like ImageNet. The weight of the CNN to train with is initialized with the weight of the pre-trained CNN model with the same architecture. Several or all layers of the network of the CNN are trained using a new dataset. When the difference between the source and the target is very significant, transfer learning mode of fine tuning is preferable. Gao et al. (2018) fine tuned all network layers of the pre-trained CNN model for the classification of interstitial lung diseases.

In this thesis work, the two main transfer learning mode based on feature extraction and fine tuning are compared with training from scratch using five selected pre-trained CNN models for breast mass classification and skin lesion classification. The next sections include the state-of-the-art in breast mass classification and skin lesion classification, materials for the thesis work, proposed methods, results, discussion, conclusion and acknowledgement.

2. State of the art

2.1. Breast Mass Classification

Some studies have demonstrated the use of automated diagnosis in the classification of breast mass. Their methods have differentiated normal tissue from masses either as main task or as an initial step in an automatic diagnosis. The studies were done with machine learning which include support vector machine (SVM) and deep learning. Their work were focused on extracting the region of interest (ROI) and then the classification into normal and benign/malignant mass.

Bruno et al. (2016) used local binary pattern (LBP) operators with curvelet coefficients to distinguish nor-

mal tissue and malignant masses, and also malignant from benign masses. Using different classifiers, the best performance was obtained by a SVM with polynomial kernel with 91% accuracy for Digital Database for Screening Mammography (DDSM) dataset.

Narváez et al. (2017) used an automatic Breast Imaging Reporting and Data System (BI-RADS) characterization of masses contained in a ROI to learn relevant radiological characteristics from various multiscale decomposition of the visual information, zernike polynomials and curvelet bases that are optimally fused by a multiple kernel learning (MKL). Their method first check the presence of masses in the ROI by training a conventional SVM classifier. When a ROI is recognized as a mass, it feeds a bank of SVM binary classifiers which selects the kind of shape from the Breast Imaging Reporting and Data System (BI-RADS) terms. The margin of the mass and density are then set by joining together the BI-RADS labels of the five most similar shapes in the database. 980 and 216 ROIs extracted from the DDSM and INbreast database respectively were used in the evaluation. Masses were detected with 96.2% sensitivity and 93.1% specificity.

In the work of Guan and Loew (2017), three methods which include CNN from scratch, transfer learning mode of feature extraction and fine tuning using VGG16 model were used to classify ROI into normal tissue and breast mass. Their best performance was on fine tuning with a validation accuracy of 91.3%.

Suzuki et al. (2016) used transfer learning mode of fine tuning with AlexNet model to classify ROI into normal and mass. Their result reported 89.9% sensitivity for mass detection.

The above mentioned studies all used ROI in the classification between normal and breast mass. In this master thesis work, breast mass classification on ROI and whole mammogram image will be studied.

2.2. Skin Lesion Classification

Several studies have been done in the automatic classification of skin lesion. Most of the studies was done towards the detection of melanoma, the most dangerous type of skin cancer.

In the work done by Joseph and Panicker (2016), they used image processing techniques which include hair detection, lesion segmentation, feature extraction and support vector machine classifier for the classification of skin lesion images. They used a two-step classification which first classifies the lesion into normal or abnormal mole. Then in the second step of classification, the abnormal mole was further classified into atypical or melanoma mole. Their method was evaluated with ph2 dataset. Their performance was 91.5% accuracy in the first step classification and 93.5% accuracy in the second step classification.

Quang et al. (2017) fine tuned with VGG16 model for the classification of skin lesion images into two categories. First task is classification into benign and malignant skin lesion and second task is classification into melanocytic and non-melanocytic lesions using ISIC 2017 skin lesion dataset. Their best performance was 76.3% and 86.9% in the first and second task respectively using the area under curve metric.

In the work of Hosny et al. (2018), fine tuning with AlexNet model was used for the classification of skin lesion into melanoma, common nevus and atypical nevus. The last layer of the AlexNet model was replaced for the classification into three classes. Their method was trained and tested with the ph2 dataset and performance was 98.6% accuracy.

Kaymak et al. (2018) using HAM10000 dataset that consists of seven types of skin lesions, fine-tuned an AlexNet model for skin lesion classification. They obtained accuracy of 78% for classification into melanocytic and non-melanocytic lesions, 84% accuracy for classification into melanoma and nevus lesions, and 58% accuracy for classification into non-melanocytic malignant and benign lesions. In their work, they did two class classification even though the dataset consist of seven types of skin lesion.

The mentioned studies were done for two or three class skin lesion classification. However, in this thesis work, seven class classification is carried out using the HAM10000 dataset that consists of seven types of skin lesion.

3. Materials

3.1. INbreast Dataset

The INbreast dataset (Moreira et al., 2012) was acquired at the breast centre located in a university hospital, Hospital de Sao Joao, Breast Centre, Porto Portugal. The images were acquired between 2008 and 2010 using the acquisition equipment MammoNovation Siemens full field digital mammography (FFDM) with a solid-state detector of amorphous selenium and 14-bit contrast selenium. The image matrix is 3328×4084 or 2560×3328 pixels which depends on the compression plate used in the acquisition. The images were saved in the DICOM format and all confidential medical information were removed from the DICOM file. The images are from screening, diagnostic and follow-up cases. The dataset contain normal mammograms, mammogram with masses, mammograms with calcifications, asymmetries, architectural distortions and images with multiple findings. In total, the dataset contains 410 images of which 107 images have mass and 303 images does not have mass. A mass which is the focus on using this dataset is defined by the BI-RADS as a three-dimensional structure demonstrating convex outward borders, usually evident on two orthogonal views (Sickles et al., 2013).

The main characteristic of this dataset is the groundtruth annotation. Most of the datasets usually

give a circle around the region of interest (ROI) but INbreast dataset have annotations that were made by a specialist in the field and validated by another specialist.

For each image, there is a binary mask to remove the black background from the mammogram image. In addition, the images with mass also have a binary groundtruth image.

Figure 6: Mammogram image with corresponding mask and groundtruth

The INbreast dataset contains 107 mammogram images with mass and 303 mammogram images without mass. In the 107 mammogram images that contain mass, some images have more than one mass in them. In total, there are 115 masses. The total mammogram images were divided into two, mammogram images with mass and without mass. The mammogram images with mass were further divided into three, 80% train, 10% validation and 10% test. The mammogram images without mass were also divided into three, 80% train, 10% validation and 10% test.

3.2. HAM10000 Dataset

The HAM10000 ("Human Against Machine with 10000 training images") dataset (Tschandl et al., 2018) are from the ISIC 2018 competition on skin lesion diagnosis. The lesion images were acquired with a variety of dermatoscope types from all the anatomic sites excluding the mucosa and nails. The lesion images are from a historical sample of patients presented for skin cancer screening from different institutions. The labels for the images were established by one of the following: histopathology, reflectance confocal microscopy, the lesion did not change during digital dermatoscopic follow up over two years with at least three images and consensus of at least three expert dermatologists from a single image. The distribution represents a modified "real world" setting. There are more benign lesions than malignant lesions. The total images in the dataset are 10015. The seven disease categories and their number of images are: melanoma - 1113, melanocytic nevus - 6705, basal cell carcinoma - 514, actinic keratosis - 327, benign keratosis - 1099, dermatofibroma - 115 and vascular lesion - 142.

The dataset is highly unbalanced. Melanocytic nevus constitutes about 67% of the dataset. The dataset is split into 80% train, 10% validation and 10% test. Distribution of the skin lesion in the dataset is shown in Figure 8.

4. Methods

4.1. Preparation of Mammogram Image Patches

To generate mammogram image patches without mass (negative mammogram image patches), images of size 454×454 were cropped from mammogram images without masses with their respective binary mask. The binary mask is used to remove the black background in the mammogram image when cropping negative mammogram image patches. Multiple patches were generated from a single image. As a result of that, many negative mammogram image patches were generated altogether. Figure 9 illustrates how negative mammogram image patches were generated.

To generate mammogram image patches with mass (positive mammogram image patches), images of size 454×454 were cropped from mammogram images with masses with their respective binary groundtruth. The binary groundtruth is used to focus on the region of interest in the mammogram image which is the mass. Centre, left and right corners of masses were extracted. Then the positive mammogram image patches were rotated by 90, 180 and 270 degrees in order to increase their number. Figure 10 illustrates how positive mammogram image patches were generated.

In total, 2000 training, 150 validation and 148 test mammogram image patches were used in this work. The two class of patches were balanced in all sets. This is as a result of augmentation done for the positive mammogram image patches to match up with the negative ones.

4.2. Base Model Architectures

Five selected base model architectures were used and they include VGG19, ResNet50, InceptionResnetV2, InceptionV3 and NASNetLarge. They are pre-trained on ImageNet database which are used in the ImageNet Large-Scale Recognition Challenge (ILSVRC).

- VGG19: Simonyan and Zisserman (2014) created a 19-layer network which consists of 16 convolution and 3 fully-connected CNN layer. The model used 3×3 filters with stride and pad of 1 with 2×2 max-pooling layers with stride 2. The VGG19 is a deeper CNN with more layers than VGG16. It uses small 3×3 filters in all convolutional layers in order to reduce the number of parameters. The VGG19 model have a total of 143,667,240 parameters and a size of 549 MB.
- ResNet50: ResNet50 is a pre-trained model which is highly useful in residual neural networks (He et al., 2016). The network learns the residuals of the input layer. Each block consists of a series of layers and a connection adding the input of the block to its output. The gradient signals can travel back directly to early layers through the various small connections and helps to solve the problem

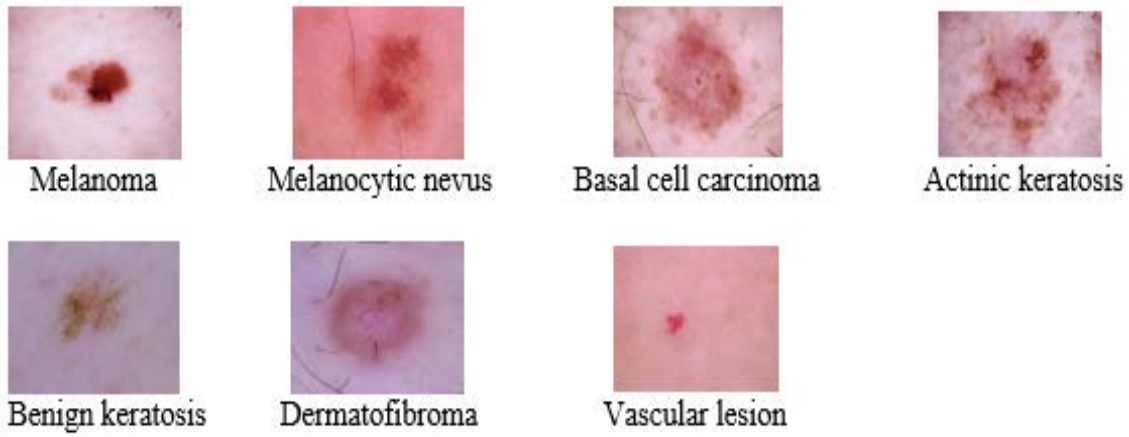


Figure 7: Typical example of the various skin lesion

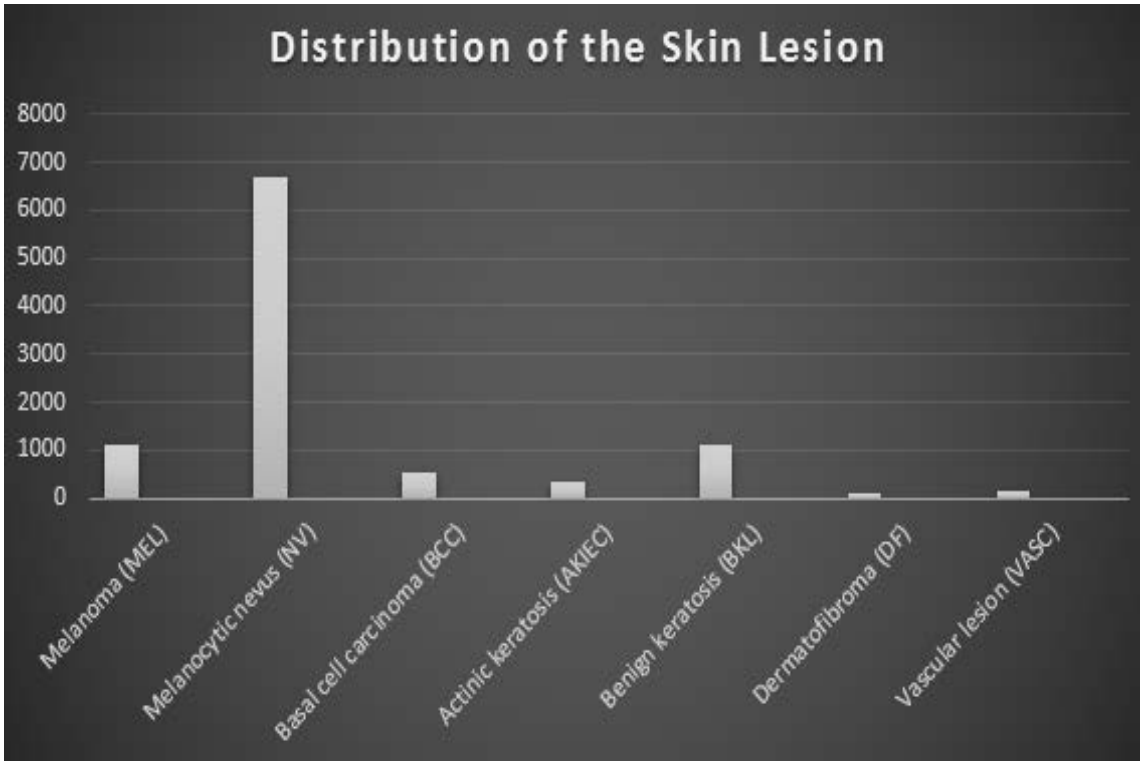


Figure 8: Distribution of the skin lesion

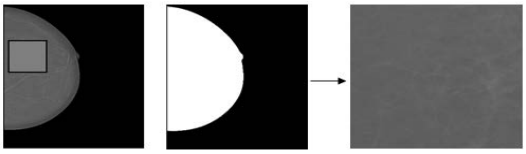


Figure 9: Creation of negative mammogram image patches

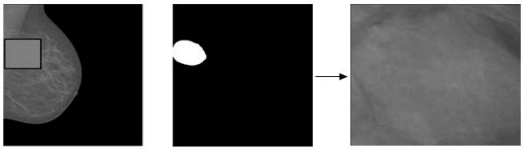


Figure 10: Creation of positive mammogram image patches

of vanishing gradients. Vanishing gradients arises when gradient signals from the error function decreases rapidly as they are back propagated to earlier layers. Some studies have shown that training residual networks are easier than a CNN. ResNet50 model have a total of 25,636,712 parameters and a size of 98 MB.

- InceptionResNetV2: InceptionResNetV2 uses residual connections that add the output of the convolution operation of the inception module to the input (Szegedy et al., 2017). The InceptionResNetV2 model have a total of 55,873,736 parameters and a size of 215 MB.
- InceptionV3: InceptionV3 is an improved version of InceptionV2 and have additions which include: RMSProp optimizer, factored 7×7 convolutions, batch normalization in the auxillary classifiers and label smoothing (Szegedy et al., 2016). The label smoothing is a kind of regularizing component that when added to the loss formula prevents the network from learning more of a particular class and prevents over fitting. The InceptionV3 model have a total of 23,851,784 parameters and a size of 92 MB.
- NASNetLarge: NASNetLarge is based on reinforcement learning algorithms (Zoph et al., 2018). It is capable of producing small-scale networks. The NASNetLarge model have a total of 88,949,818 parameters and a size of 343 MB.

4.3. Classifier on Top of the Base Model

The fully connected layers on top of the five selected models were removed. A new classifier on top of the base models was designed. The classifier on top of the base models for breast mass classification and skin lesion classification are different. This is due to the difference in the number of classes and their distribution in the two training sets. In this thesis work, a novel neural network technique in which the classifier is concatenated was adopted for breast mass classification. Compared with Nguyen et al. (2018) that used three network architectures (InceptionV3, ResNet512 and InceptionResNetV2) for feature concatenation in microscopic image classification, concatenation of classifier in this work is done with the same network architecture. This approach gives a superior performance for training set that have balanced class distribution. In the skin lesion classification, concatenation of classifier was not applied because the approach gives a lower performance than without concatenation of classifier. Also for the skin lesion classification, more layers in the classifier tend to reduce the performance and so less layers were adopted. The classifier on top of the base models for breast mass classification and skin lesion classification are shown in Figure 11 and 12 respectively.

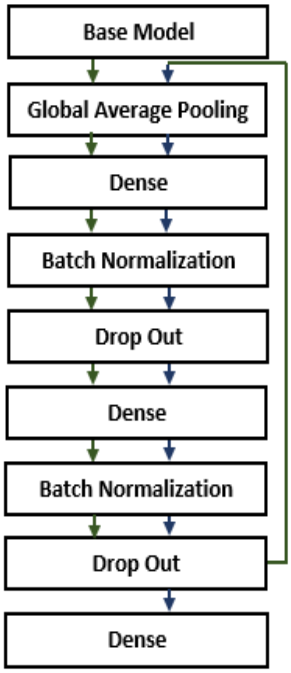


Figure 11: Classifier for breast mass. The green arrows illustrates before concatenation and the blue arrows illustrates after concatenation

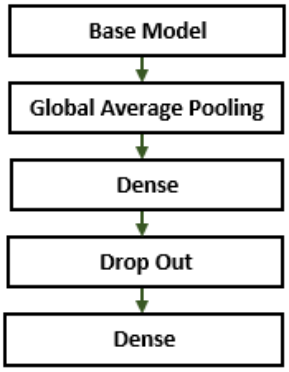


Figure 12: Classifier for skin lesion

Global average pooling was proposed by Lin et al. (2013) to replace the fully connected layers and it com-

putes the average value of all the elements in the feature map. It is used to generate one feature map for each corresponding category of the classification task in the last convolutional layer. The global average pooling is more native to the convolution structure by enforcing correspondences between feature maps and categories. This makes feature maps to be easily interpreted as categories confidence maps. It also prevents overfitting and is more efficient to spatial translations of the input.

The dense layer also called a fully connected layer connects every node in one layer to every node in another layer. The number of nodes in the first and second dense layers in the classifier for breast mass is 1024 and 128 respectively. The number of nodes in the first dense layer in the classifier for skin lesion is 1024. The last dense layer which is the output layer performs classification based on the features extracted by the previous layers and the number of nodes in it corresponds to the number of classes to be predicted, two for breast mass classification and seven for skin lesion classification.

The dense layer also contains activation function. The activation function adds the nonlinear factors so that redundant data are removed while preserving features. The activation function also retains active neuron feature and maps out these features by nonlinear functions and this is essential for the neural network to solve the complex nonlinear problems. ReLu activation function was used in the dense layer that does not do the prediction. ReLu's output is zero and maximum of input. Its output equals to input when input value is non-negative and this alleviate the gradient vanishing and exploding problems. ReLu activation function do not have boundaries. Its range is from zero to infinity and so is not good for prediction in the output layer. Softmax activation function was used in the last dense layer that do the prediction because its range is from zero to one and can regulate the output values (Hinton and Salakhutdinov, 2009).

Batch Normalization was proposed by Ioffe and Szegedy (2015). Batch normalization reduces internal covariate shift and this improves the training of deep neural nets. Internal covariate shift is the change in the distribution of network activations due to change in network parameters during training. Batch normalization reduces the internal covariate shift by using a normalization step which fixes the means and variances of layer inputs. It has a good effect on the gradient flow through the network by decreasing the influence of gradients on the scale of the parameters. It also regularizes the model. Batch normalization when combined with ReLu activation function achieves a great performance in feature extraction. Some state-of-the-art "non-plain" CNN structures, He et al. (2016) and Huang et al. (2017) use batch normalization with ReLu as a feature normalize and nonlinear transform operator. However, the combination of batch normalization and ReLu will extract more active features from previous feature maps and this improves efficiency but makes feature maps sparse. This feature maps that are sparse will reduce the performance in dataset that are greatly unbalanced. ReLu activation is used in the two classifiers and so adding batch normalization will affect the classification in dataset that is greatly unbalanced. Batch normalization was added only in the classifier for breast mass because the training set is balanced in distribution. The training set of skin lesion is greatly unbalanced in distribution and batch normalization was not used in the classifier for skin lesion.

Dropout was proposed by Srivastava et al. (2014). It is a simple regularization technique to solve overfitting. Overfitting occurs when the model includes more terms or complicated approaches than what is needed. This mostly occur when limited data in the training set is used on deep network. It leads to a high accuracy on the training set but a low accuracy on the test set. Dropout prevents overfitting on the training set and improves performance by combining exponentially many different neural network models efficiently. It chooses units to drop out randomly and removes them from the layer temporarily. The dropout rate in both the classifier for breast mass and skin lesion is 0.5.

4.4. Training Approaches

Three training approaches were adopted and was done using the five selected network architectures that were used in ImageNet classification. They include transfer learning mode of feature extraction, transfer learning mode of fine tuning and training from scratch.

- Transfer Learning Mode of Feature Extraction: In transfer learning mode of feature extraction, all the layers in the base model were frozen and only the classifier was trained. The weight is initialized to ImageNet.
- Transfer Learning Mode of Fine Tuning: In transfer learning mode of fine tuning, all the layers were trained and weight initialization to ImageNet except for NASNetLarge model in training for mammogram image patches in which the first 600 layers of the base model were frozen. The first 600 layers of the base model were frozen when using the NASNetLarge model for training of mammogram image patches because concatenation of classifier is applied. This takes more memory and made NASNetLarge model to run out of memory when all layers were to be trained.
- Training from Scratch: In training from scratch, all the layers were trained except for NASNetLarge model in training for mammogram image patches in which the first 600 layers of the base model were frozen. The same reason given above for freezing the first 600 layers of the base model of NASNetLarge model also applies here. There is no weight initialization to ImageNet.

The number of total and trainable parameters for the five selected models in transfer learning mode of feature extraction, transfer learning mode of fine tuning and training from scratch for breast mass and skin lesion classification are shown in Table 1 and Table 2 respectively.

4.5. Data Augmentation During Training

Deep neural networks usually require large number of training data to achieve good performance. Data augmentation can increase the size of the training data by generating new data from the original input data. Data augmentation also generates various versions of training samples because masses and lesions can appear in various orientations. Data augmentation during training helps to minimize disk space to store the augmented images. After the training, the augmented data are removed from the system automatically.

For the mammogram image patches, each original image was flipped horizontally, rotated by 20 degrees, zoomed by 0.2, sheared by 0.2, shifted the width and height by 0.1. Each original image of the mammogram patch was thus augmented to six images.

For the skin lesion images, each original image was flipped horizontally, rotated by 180 degrees, zoomed by 0.2, shifted the width and height by 0.1. Each original image of the skin lesion was thus augmented to five images.

4.6. Optimizer

Optimizer update the weight parameters to minimize the loss function. Three optimizers were used in this work. They include Adam (Kingma and Ba, 2014), RMSProp (Tieleman and Hinton, 2012) and Stochastic gradient descent (SGD). Adaptive gradient optimizers which include Adam and RMSProp have a rapid training time and take less time to reach convergence. Adaptive gradient optimizers are also ideal for small dataset. Non-adaptive optimizers which include SGD take more time during training.

In breast mass classification, Adam was used in all training approaches. For skin lesion, RMSProp was used in transfer learning mode of feature extraction and SGD with momentum was used in transfer learning mode of fine tuning and training from scratch. Momentum is a method that helps move SGD in the right direction and dampens oscillations (Qian, 1999). It does this by adding a fraction γ of the update vector of the previous time step to the current update vector. Momentum helps to gain faster convergence and reduced oscillation. The term for the momentum γ is set to 0.9.

4.7. Learning Rate and other Hyperparameters

Learning rate is a hyperparameter that controls how the weights of the network are adjusted with respect to the loss gradient. A learning rate that is too small leads to a slow convergence while a learning rate that is too large can limit convergence and make the loss function to fluctuate around the minimum. The learning rate for the breast mass classification is 0.00001 and was increased to 0.001 for skin lesion classification because of the large training set. Other hyperparameters were kept at their default value.

4.8. Batch Size and Epoch

Batch size is the number of samples to pass through before the model is updated. Epoch is the number of times for the learning algorithm to pass through the training set. The batch size for all the models is 16 except NASNetLarge model that is 8 due to large memory. The number of epoch is set to 200 for all training.

4.9. Save Best Weight During Training and Early Stopping

In training, the weight of the improved validation accuracy is saved and updated. After 20 epochs, if there is no further improvement in the validation accuracy, the training is automatically stopped. This is helpful because the last epoch may not give the best validation accuracy to be updated in the weight.

4.10. Performance Evaluation

The performance evaluation on the test set for the three training approaches using the five selected models were done with the accuracy metrics. The best performance was further evaluated using the normalized confusion matrix and area under the receiver operating characteristic curve (AUC).

Accuracy is the ratio of the number of correct predictions to the total number of samples.

Normalized confusion matrix gives a matrix as output and describes the complete performance of the model by showing the percentage of the correct and incorrect predictions.

Area under the receiver operating characteristic curve (AUC) is the probability that the classifier will rank a randomly chosen sample higher than a randomly chosen of other samples.

4.11. Hardware and Software Environment

The work was developed with Python programming language using Keras with Tensorflow backend. Keras library have the five selected models used in this work which are pre-trained on ImageNet. The work was done with a remote machine that have the following features: CPU of 1 with 8 hyperthreaded core, 256 GB RAM and GPU of 2 TitanXp with 12GB onboard.

Model	Total Parameters	Trainable Parameters		
		Transfer Learning		Train from Scratch
		Feature Extraction	Fine Tune	
VGG19	20,686,786	660,098	20,684,482	20,684,482
ResNet50	25,826,050	2,236,034	25,770,626	25,770,626
InceptionResNetV2	56,049,762	1,710,722	55,986,914	55,986,914
InceptionV3	24,041,122	2,236,034	24,004,386	24,004,386
NASNetLarge	89,190,740	4,271,618	73,976,834	73,976,834

Table 1: Total and Trainable Parameters for Breast Mass Classification

Model	Total Parameters	Trainable Parameters		
		Transfer Learning		Train from Scratch
		Feature Extraction	Fine Tune	
VGG19	20,556,871	532,487	20,556,871	20,556,871
ResNet50	25,693,063	2,105,351	25,639,943	25,639,943
InceptionResNetV2	55,917,799	1,581,063	55,857,255	55,857,255
InceptionV3	23,908,135	2,105,351	23,873,703	23,873,703
NASNetLarge	89,053,785	4,136,967	88,857,117	88,857,117

Table 2: Total and Trainable Parameters for Skin Lesion Classification

5. Results

5.1. Breast Mass Classification

In all the three training approaches using the five selected models, the performance using accuracy metrics was evaluated on the test set of the mammogram image patches. The performance is shown in Table 3.

Further performance evaluation was done with the best performance in Table 3 using normalized confusion matrix and receiver operating characteristic (ROC) curve. This is shown in Figure 13 and 14.

Using the weight of the best model performance, prediction was done on a whole mammogram image. In this, a frame work was designed in which the binary mask was used to remove the black background from the whole mammogram image so that only the breast area will be left. The breast area is then forwarded to the classifier to predict whether there is breast mass or not. The framework was used to predict on the test set of the whole mammogram images. The accuracy was 0.72 which is below 0.96 the model gave on the test set of the image patches.

5.2. Skin Lesion Classification

In all the three training approaches using the five selected models, the performance using accuracy metrics was evaluated on the test set of the skin lesion images. The performance are shown in Table 4.

Further performance evaluation was done with the best performance in Table 4 using normalized confusion matrix and receiver operating characteristic (ROC) curve. This is shown in Figure 15 and 16.

6. Discussion

6.1. Breast Mass Classification

Three training approaches namely transfer learning mode of feature extraction, transfer learning mode of fine tuning and training from scratch were used in the classification of breast mass using five selected models. The training was done on 12,000 augmented mammogram image patches and tested on 148 mammogram image patches. Transfer learning mode of feature extraction have the least performance in all models. This is because mammogram image differ so much from the natural images in the ImageNet dataset. Transfer learning mode of feature extraction is helpful when the training set is small as it helps to control overfitting. In the three training approaches, transfer learning mode of fine tuning gave the best performance in four out of five models. The exception was seen in NASNetLarge in which the first 600 layers of the base model was frozen and the remaining layers was trained for both in fine tuning and training from scratch. The reason for this exception from the other four model performance could be that sometimes in training, due to memory issues the result to expect may not be seen. When many people use the GPU at the same time, the result may vary from what to expect. The best result overall was in transfer learning mode of fine tuning with ResNet50 giving an accuracy of 0.96 on the test set of mammogram image patches. With four models having better performance in transfer learning mode of fine tuning than training from scratch, it can be seen that transfer learning mode of fine tuning is the ideal to be adopted in breast mass classification with mammography image modality when the training set is large.

Model	Transfer Learning		Train from Scratch
	Feature Extraction	Fine Tune	
VGG19	0.78	0.95	0.93
ResNet50	0.79	**0.96**	0.91
InceptionResNetV2	0.76	0.93	0.91
InceptionV3	0.82	0.93	0.91
NASNetLarge	0.78	0.82	0.86

Table 3: Performance using accuracy metrics on the test set of mammogram image patches

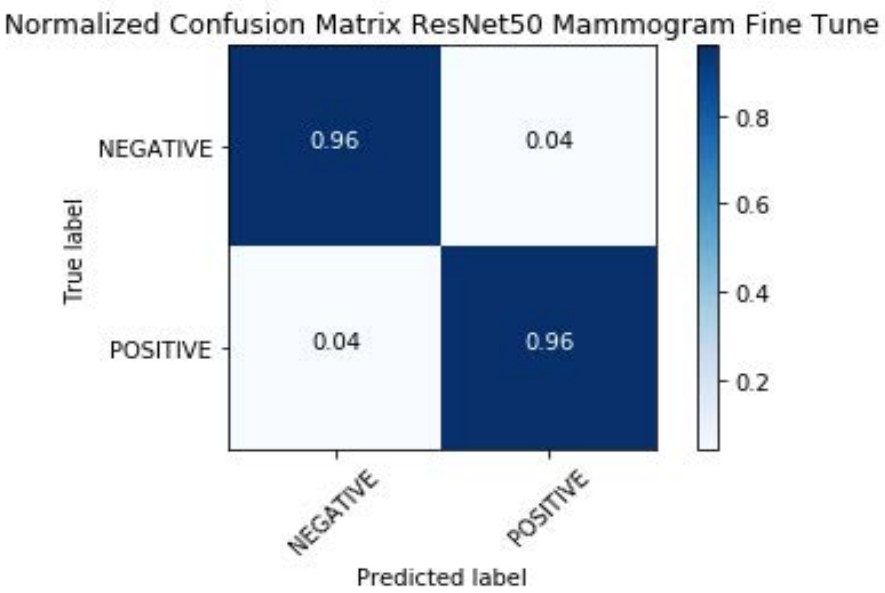


Figure 13: Normalized Confusion Matrix with ResNet50 on the test set of mammogram image patches

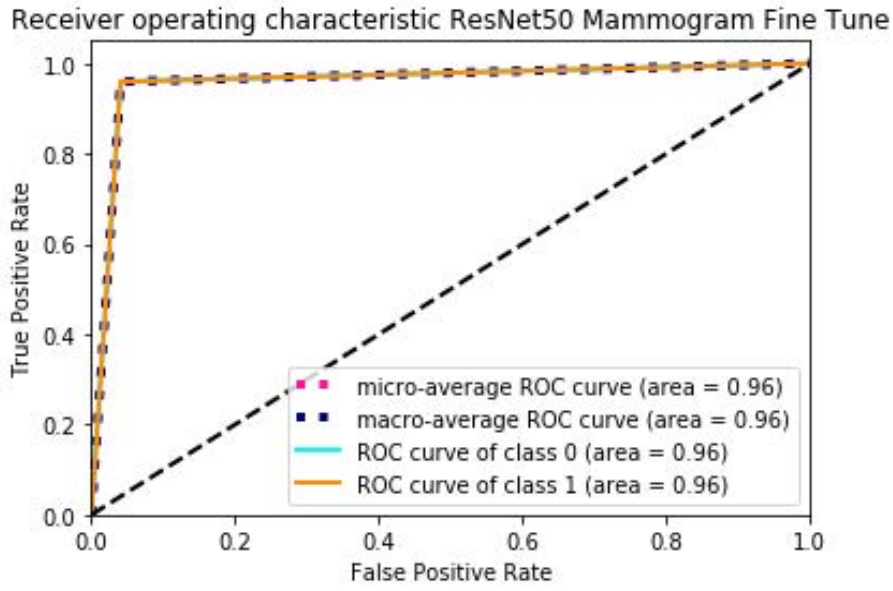


Figure 14: ROC curve with ResNet50 on the test set of mammogram image patches

Model	Transfer Learning		Train from Scratch
	Feature Extraction	Fine Tune	
VGG19	0.68	0.80	0.81
ResNet50	0.68	0.79	0.81
InceptionResNetV2	0.70	0.73	0.76
InceptionV3	0.72	**0.83**	0.78
NASNetLarge	0.71	0.77	0.81

Table 4: Performance using accuracy metrics on the test set of skin lesion images

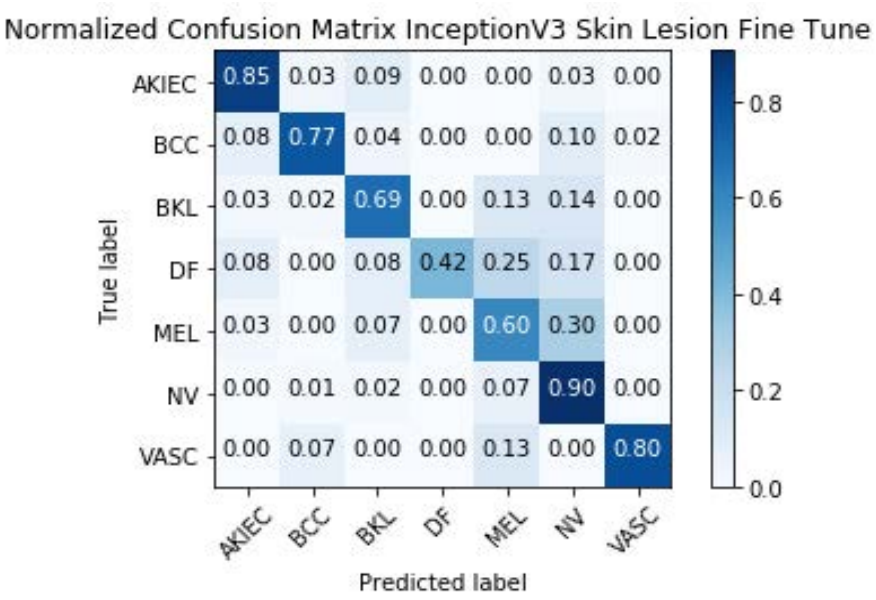


Figure 15: Normalized Confusion Matrix with InceptionV3 on the test set of skin lesion images

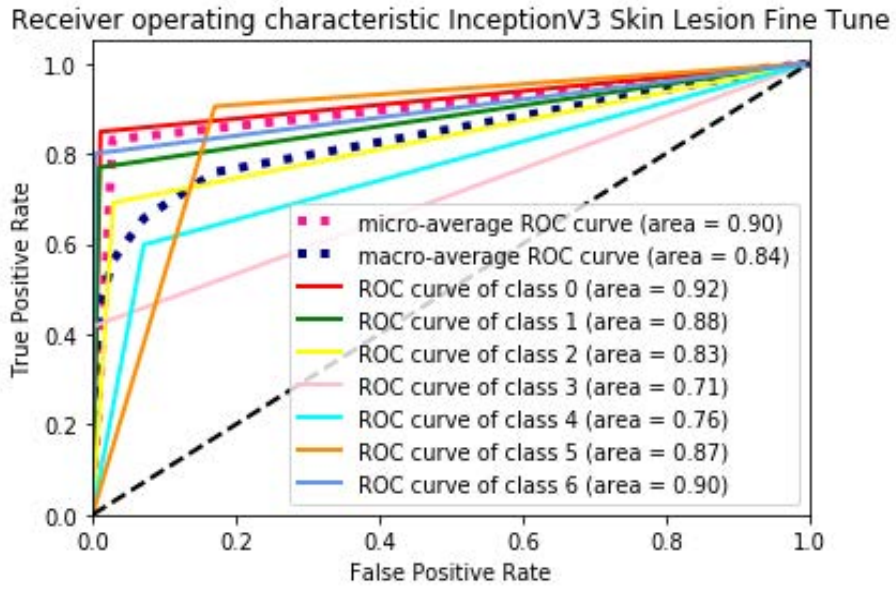


Figure 16: ROC curve with InceptionV3 on the test set of skin lesion images

The false negatives and false positives on the test set of mammogram image patches were analyzed. The false negatives were due to the breast mass being very small and not clearly visible in the mammogram image patch and so the classifier see it as negative even in actual sense it contains breast mass. The false positives were due to the presence of dense tissues in the mammogram image patch. The classifier see the dense tissue as masses. Two samples of false negatives and false positives of mammogram image patches are shown in Figure 17 and Figure 18 respectively. Tomosynthesis a new technique of mammography in 3D helps to solve the problem of the dense tissue and makes it not to be clearly visible in the 3D tomosynthesis image.

The best model performance was used to predict on the test set of whole mammogram images because in clinical practice whole mammogram images are used and not the mammogram image patches. However, predicting on a whole mammogram image is more difficult than the mammogram image patch because of the expanded size and feature space. The accuracy was 0.72 on the test set of whole mammogram images which is below 0.96 the model gave on the test set of mammogram image patches. The training was done on the mammogram image patches in which the breast mass appear bigger and when predicting on whole mammogram image in which the breast mass appear smaller, the classifier could not recognize some whole mammogram image in which the breast mass is small. However, in the whole mammogram image in which the breast mass is bigger, the classifier predicted accurately of breast mass in them. Also the presence of dense tissues overlapping with mass was also a problem in predicting a false negative in whole mammogram image. In addition, presence of dense tissue also make the classifier to predict a false positive in whole mammogram image in which it see the dense tissue as a mass. Two samples of false negatives and false positives on whole mammogram images are shown in Figure 19 and Figure 20 respectively.

Tomosynthesis a new technique of mammography in 3D in which the overlay and presence of dense tissues is diminished will play a vital role in effective classification of breast mass in both mammogram image patches and whole mammogram images.

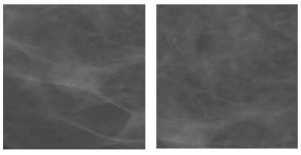


Figure 17: False negatives on mammogram image patch

6.2. Skin Lesion Classification

Transfer learning mode of feature extraction, transfer learning mode of fine tuning and training from scratch were the three training approaches used in the classification of skin lesions using five selected models. 40,050 augmented skin lesion images were used for the training and tested on 1005 skin lesion images. The least performance in all models were seen in the transfer learning mode of feature extraction. This is because the skin lesions have features that are different from the natural images in ImageNet. The best performance in four out of five models were seen in training from scratch. The exception was seen in InceptionV3 in which the transfer learning mode of fine tuning gave better performance than training from scratch and that gave the overall best performance with accuracy of 0.83. Like stated before, many factors like GPU memory usage at a time can make performance to vary. The difference between the performance in training from scratch and transfer learning mode of fine tuning were not so significant. Even though the best performance was seen in transfer learning mode of fine tuning with InceptionV3, it can be observed that training from scratch is ideal to be adopted in skin lesion classification when the training set is large considering their general performance when compared to the transfer learning mode of fine tuning.

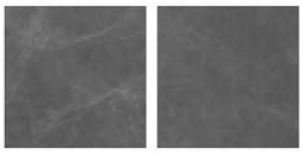


Figure 18: False positives on mammogram image patch

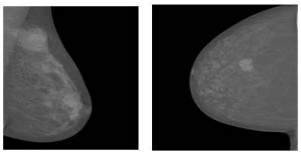


Figure 19: False negatives on whole mammogram image

The training set were greatly unbalanced with melanocytic nevus (NV) consisting of 67% of the entire set. The effect of the great unbalance in the training set was seen only in the transfer leaning mode of feature extraction because of the large data. It did not have effect on the model performance with the transfer learning mode of fine tuning and training from scratch because some model have better individual performance for other skin lesions than the melanocytic nevus. Global average pooling in the classifier which generates one feature map for each corresponding category of the classification task in the last convolutional layer helped to minimize the great unbalance in the training set.

Dermatofibroma gave low performance generally in all the predictions. This type of skin lesion are small harmless growths that appear on the skin. This type of skin lesion vary so much in color, size and features

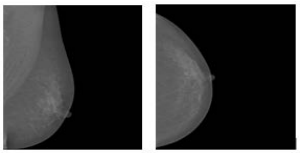


Figure 20: False positives on whole mammogram image

among them. This variation has made them to be difficult for learning and accurate prediction on the test set. Increase in the number of their training images can boost performance.

7. Conclusions

In transfer learning, the information obtained by a trained model is re-used on another task. Five selected architectures were adopted in this work which include VGG19, ResNet50, InceptionResNetv2, InceptionV3 and NASNetLarge. These architectures were selected because they are available in the Keras library and the architectures were deep enough that the three training approaches are suitable for it. Deeper architectures have recently shown high performance in various tasks of computer vision but their significant performance in medical imaging have not thoroughly been investigated in all medical imaging modalities.

The two modes of transfer learning which include feature extraction and fine tuning have situations in which one is preferable than the other. Like when the training set is small, transfer learning mode of feature extraction seems to be the ideal one because it helps to control overfitting but when the training set is large, transfer learning mode of fine tuning appears to be the better one to be adopted. Training from scratch also have good performance on large training set. From the observations in this work, there is not much significant difference between the performance in transfer learning mode of fine tuning and training from scratch. In breast mass classification using the INbreast dataset, transfer learning mode of fine tuning gave better performance than training from scratch in four out of five models while in skin lesion classification using the HAM10000 dataset, training from scratch gave better performance in four out of five models. In overall, the best performance for both classification were in transfer learning mode of fine tuning. The performance show that the training approach to be adopted between transfer learning mode of fine tuning and training from scratch when the amount of labeled data is large depends on the application and medical imaging modality.

In addition, batch normalization and the number of layers in the classifier play a vital role in image classification performance. From the findings in this work, training set in breast mass classification which have balanced classes got better performance with batch normalization and more number of layers in the classifier while the training set in skin lesion classification in which the classes are highly unbalanced got better performance when batch normalization is removed and less number of layers in the classifier.

The performance reported for the five selected models in the three training approaches in both classification can still be improved by preprocessing of the labeled data before training. Preprocessing of the data was not done in this work.

Due to limited time, only two medical imaging modalities were covered which include mammography and dermoscopy. Future work will be to study the three training approaches in other medical imaging modalities like MRI, CT, and Ultrasound using large dataset.

8. Acknowledgments

I would like to thank to my supervisor Dr Mario Molinara for his advice and suggestions throughout my master thesis.

I would also like to thank my Italian friends Antonio, Paola, Alessandro, Daniele, Marina etc. for the wonderful time shared with them during my master thesis.

Finally, I would like to thank my parents and siblings for their encouragement in my master thesis period.

References

Abdel-Razeq, H., Al-Omari, A., Zahran, F., Arun, B., 2018. Germline brca1/brca2 mutations among high risk breast cancer patients in jordan. BMC cancer 18, 152.

Argenziano, G., Fabbrocini, G., Carli, P., De Giorgi, V., Sammarco, E., Delfino, M., 1998. Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: comparison of the abcd rule of dermatoscopy and a new 7-point checklist based on pattern analysis. Archives of dermatology 134, 1563–1570.

Baldwin, L., Dunn, J., 2013. Global controversies and advances in skin cancer. Asian Pacific Journal of Cancer Prevention 14, 2155–2157.

Bar, Y., Diamant, I., Wolf, L., Lieberman, S., Konen, E., Greenspan, H., 2015. Chest pathology detection using deep learning with non-medical training, in: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 294–297.

Beral, V., Collaborators, M.W.S., et al., 2007. Ovarian cancer and hormone replacement therapy in the million women study. The Lancet 369, 1703–1710.

Blum, A., Rassner, G., Garbe, C., 2003. Modified abc-point list of dermoscopy: a simplified and highly accurate dermoscopic algorithm for the diagnosis of cutaneous melanocytic lesions. Journal of the American Academy of Dermatology 48, 672–678.

Boyd, N., Byng, J., Jong, R., Fishell, E., Little, L., Miller, A., Lockwood, G., Tritchler, D., Yaffe, M.J., 1995. Quantitative classification of mammographic densities and breast cancer risk: results from the canadian national breast screening study. JNCI: Journal of the National Cancer Institute 87, 670–675.

Boyd, N.F., Dite, G.S., Stone, J., Gunasekara, A., English, D.R., McCredie, M.R., Giles, G.G., Tritchler, D., Chiarelli, A., Yaffe, M.J., et al., 2002. Heritability of mammographic density, a risk factor for breast cancer. New England Journal of Medicine 347, 886–894.

Boyer, B., Canale, S., Arfi-Rouche, J., Monzani, Q., Khaled, W., Balleyguier, C., 2013. Variability and errors when applying the birads mammography classification. European journal of radiology 82, 388–397.

Bruno, D.O.T., do Nascimento, M.Z., Ramos, R.P., Batista, V.R., Neves, L.A., Martins, A.S., 2016. Lbp operators on curvelet coefficients as an algorithm to describe texture in breast cancer tissues. Expert Systems with Applications 55, 329–340.

Cascinelli, N., Ferrario, M., Tonelli, T., Leo, E., 1987. A possible new tool for clinical diagnosis of melanoma: the computer. Journal of the American Academy of Dermatology 16, 361–367.

Clemmesen, J., 1948. Carcinoma of the breast. results from statistical research. Br J Radiol 21, 583–590.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection, in: international Conference on computer vision & Pattern Recognition (CVPR'05), IEEE Computer Society. pp. 886–893.

D'Alessandro, B., Dhawan, A.P., 2012. 3-d volume reconstruction of skin lesions for melanin and blood volume estimation and lesion severity analysis. IEEE transactions on medical imaging 31, 2083–2092.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee. pp. 248–255.

Dhawan, A.P., Gordon, R., Rangayyan, R.M., 1984. Nevoscopy: three-dimensional computed tomography of nevi and melanomas in situ by transillumination. IEEE transactions on medical imaging 3, 54–61.

Easton, D.F., Ford, D., Bishop, D.T., 1995. Breast and ovarian cancer incidence in brca1-mutation carriers. breast cancer linkage consortium. American journal of human genetics 56, 265.

Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D., Forman, D., Bray, F., 2013. Globocan 2012 v1. 0. Cancer incidence and mortality worldwide: IARC CancerBase 11.

Ganapathy, N., Swaminathan, R., Deserno, T.M., 2018. Deep learning on 1-d biosignals: a taxonomy-based survey. Yearbook of medical informatics 27, 098–109.

Gao, M., Bagci, U., Lu, L., Wu, A., Buty, M., Shin, H.C., Roth, H., Papadakis, G.Z., Depeursinge, A., Summers, R.M., et al., 2018. Holistic classification of ct attenuation patterns for interstitial lung diseases via deep convolutional neural networks. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 6, 1–6.

Ghoussaini, M., Pharoah, P.D., Easton, D.F., 2013. Inherited genetic susceptibility to breast cancer: the beginning of the end or the end of the beginning? The American journal of pathology 183, 1038–1051.

Grimm, L.J., Ghate, S.V., Yoon, S.C., Kuzmiak, C.M., Kim, C., Mazurowski, M.A., 2014. Predicting error in detecting mammographic masses among radiology trainees using statistical models based on bi-rads features. Medical physics 41.

Guan, S., Loew, M., 2017. Breast cancer detection using transfer learning in convolutional neural networks, in: 2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), IEEE. pp. 1–8.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Hinton, G.E., Salakhutdinov, R.R., 2009. Replicated softmax: an undirected topic model, in: Advances in neural information processing systems, pp. 1607–1614.

Hosny, K.M., Kassem, M.A., Foaud, M.M., 2018. Skin cancer classification using deep learning and transfer learning, in: 2018 9th Cairo International Biomedical Engineering Conference (CIBEC), IEEE. pp. 90–93.

Howlader, N., Noone, A., Krapcho, M., Garshell, J., Neyman, N., Altekruse, S., Kosary, C., Yu, M., Ruhl, J., Tatalovich, Z., et al., 2013. Seer cancer statistics review, 1975–2010, national cancer institute. bethesda, md, based on november 2012 seer data submission, posted to the seer web site; 2013. seer. cancer. gov/csr/1975_2010 (Accessed June 08, 2013) .

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708.

INTERSUN, W., 2003. the global uv project: a guide and compendium. geneva, switzerland. World Health Organization .

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 .

Joseph, S., Panicker, J.R., 2016. Skin lesion analysis system for melanoma detection with an effective hair segmentation method, in: 2016 International Conference on Information Science (ICIS), IEEE. pp. 91–96.

Kalinin, A.A., Higgins, G.A., Reamaroon, N., Soroushmehr, S., Allyn-Feuer, A., Dinov, I.D., Najarian, K., Athey, B.D., 2018. Deep learning in pharmacogenomics: from gene regulation to patient stratification. Pharmacogenomics 19, 629–650.

Katzman, J.L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y., 2018. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. BMC medical research methodology 18, 24.

Kaymak, S., Esmaili, P., Serener, A., 2018. Deep learning for two-step classification of malignant pigmented skin lesions, in: 2018 14th Symposium on Neural Networks and Applications (NEUREL), IEEE. pp. 1–6.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.

Kwon, J.m., Lee, Y., Lee, Y., Lee, S., Park, J., 2018. An algorithm based on deep learning for predicting in-hospital cardiac arrest. Journal of the American Heart Association 7, e008678.

Lewis, T.C., Pizzitola, V.J., Giurescu, M.E., Eversman, W.G., Lorans, R., Robinson, K.A., Patel, B.K., 2017. Contrast-enhanced digital mammography: A single-institution experience of the first 208 cases. The breast journal 23, 67–76.

Lin, M., Chen, Q., Yan, S., 2013. Network in network. arXiv preprint arXiv:1312.4400 .

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. International journal of computer vision 60, 91–110.

Matsuda, A., Matsuda, T., Shibata, A., Katanoda, K., Sobue, T., Nishimoto, H., Group, J.C.S.R., 2014. Cancer incidence and incidence rates in japan in 2008: a study of 25 population-based cancer registries for the monitoring of cancer incidence in japan (mcij) project. Japanese journal of clinical oncology 44, 388–396.

Mayer, J., et al., 1997. Systematic review of the diagnostic accuracy of dermatoscopy in detecting malignant melanoma. Medical journal of Australia 167, 206–210.

Menzies, S.W., Ingvar, C., Crotty, K.A., McCarthy, W.H., 1996. Frequency and morphologic characteristics of invasive melanomas lacking specific surface microscopic features. Archives of Dermatology 132, 1178–1182.

Moreira, I.C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M.J., Cardoso, J.S., 2012. Inbreast: toward a full-field digital mammographic database. Academic radiology 19, 236–248.

Narváez, F., Díaz, G., Poveda, C., Romero, E., 2017. An automatic bi-rads description of mammographic masses by fusing multiresolution features. Expert Systems with Applications 74, 82–95.

Nguyen, L.D., Lin, D., Lin, Z., Cao, J., 2018. Deep cnns for microscopic image classification by exploiting transfer learning and feature concatenation, in: 2018 IEEE International Symposium on Circuits and Systems (ISCAS), IEEE. pp. 1–5.

Oliver, A., Freixenet, J., Marti, J., Perez, E., Pont, J., Denton, E.R., Zwiggelaar, R., 2010. A review of automatic mass detection and segmentation in mammographic images. Medical image analysis 14, 87–110.

Oquab, M., Bottou, L., Laptev, I., Sivic, J., 2014. Learning and transferring mid-level image representations using convolutional neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1717–1724.

Pereyra, M., Dobigeon, N., Batatia, H., Tourneret, J.Y., . Segmentation of skin lesions in 2d and 3d ultrasound images using a spatially coherent generalized rayleigh mixture model .

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L., 2018. Deep contextualized word representations. arXiv preprint arXiv:1802.05365 .

Qian, N., 1999. On the momentum term in gradient descent learning algorithms. Neural networks 12, 145–151.

Quang, N.H., et al., 2017. Automatic skin lesion analysis towards melanoma detection, in: 2017 21st Asia Pacific Symposium on Intelligent and Evolutionary Systems (IES), IEEE. pp. 106–111.

Raina, R., Madhavan, A., Ng, A.Y., 2009. Large-scale deep unsupervised learning using graphics processors, in: Proceedings of the 26th annual international conference on machine learning, ACM. pp. 873–880.

Satheesha, T., Satyanarayana, D., Prasad, M.G., Dhruve, K.D., 2017. Melanoma is skin deep: a 3d reconstruction technique for computerized dermoscopic skin lesion classification. IEEE journal of translational engineering in health and medicine 5, 1–17.

Shickel, B., Tighe, P.J., Bihorac, A., Rashidi, P., 2017. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. IEEE journal of biomedical and health informatics 22, 1589–1604.

Sickles, E., d'Orsi, C., Bassett, L., Appleton, C., Berg, W., Burnside, E., et al., 2013. Acr bi-rads® mammography. ACR BI-RADS® Atlas, Breast imaging reporting and data system 5.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research 15, 1929–1958.

Stolz, W., 1994. Abcd rule of dermatoscopy: a new practical method for early recognition of malignant melanoma. Eur. J. Dermatol. 4, 521–527.

Suzuki, S., Zhang, X., Homma, N., Ichiji, K., Sugita, N., Kawasumi, Y., Ishibashi, T., Yoshizawa, M., 2016. Mass detection using deep convolutional neural network for mammographic computer-aided diagnosis, in: 2016 55th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE), IEEE. pp. 1382–1386.

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning, in: Thirty-First AAAI Conference on Artificial Intelligence.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826.

Tang, J., Rangayyan, R.M., Xu, J., El Naqa, I., Yang, Y., 2009. Computer-aided detection and diagnosis of breast cancer with mammography: recent advances. IEEE Transactions on Information Technology in Biomedicine 13, 236–251.

Tieleman, T., Hinton, G., 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning 4, 26–31.

Ting, F.F., Tan, Y.J., Sim, K.S., 2019. Convolutional neural network improvement for breast cancer classification. Expert Systems with Applications 120, 103–115.

Tittmann, B.R., Miyasaka, C., Maeva, E., Shum, D., 2013. Fine mapping of tissue properties on excised samples of melanoma and skin without the need for histological staining. IEEE transactions on ultrasonics, ferroelectrics, and frequency control 60, 320–331.

Travis, L.B., Hill, D., Dores, G.M., Gospodarowicz, M., Van Leeuwen, F.E., Holowaty, E., Glimelius, B., Andersson, M., Pukkala, E., Lynch, C.F., et al., 2005. Cumulative absolute breast cancer risk for young women treated for hodgkin lymphoma. Journal of the National Cancer Institute 97, 1428–1437.

Tschandl, P., Rosendahl, C., Kittler, H., 2018. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data 5, 180161.

Vogt, M., Ermert, H., 2007. In vivo ultrasound biomicroscopy of skin: Spectral system characteristics and inverse filtering optimization. IEEE transactions on ultrasonics, ferroelectrics, and frequency control 54, 1551–1559.

Vranas, K.C., Jopling, J.K., Sweeney, T.E., Ramsey, M.C., Milstein, A.S., Slatore, C.G., Escobar, G.J., Liu, V.X., 2017. Identifying distinct subgroups of intensive care unit patients: a machine learning approach. Critical care medicine 45, 1607.

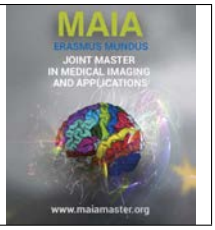
Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., Collins, N., Gregory, S., Gumbs, C., Micklem, G., et al., 1995. Identification of the breast cancer susceptibility gene brca2. Nature 378, 789.

Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., Stolcke, A., 2018. The microsoft 2017 conversational speech recognition system, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 5934–5938.

Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V., 2018. Learning transferable architectures for scalable image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8697–8710.

Medical Imaging and Applications

Master Thesis, June 2019



Improving the Detection of Autism Spectrum Disorder by Combining Structural and Functional MRI Information

Mladen Rakić, Mariano Cabezas, Arnau Oliver, Xavier Lladó

Computer Vision and Robotics Group, University of Girona, Catalonia, Spain

Abstract

Autism Spectrum Disorder (ASD) is a brain disorder, typically characterized by deficits in social communication and interaction, as well as restrictive and repetitive behaviors and interests. In this master thesis, a deep learning-based method for classification of ASD versus typical control subjects is proposed. The method is based on incorporation of both functional and structural information with the goal of maximizing classification accuracy. Functional connectivity patterns among brain regions, together with correlations of gray matter volumes among cortical parcels are used as features for functional and structural processing pipelines, respectively. The classification network is a combination of stacked autoencoders trained in an unsupervised manner and multi-layer perceptrons trained in a supervised manner. Quantitative analysis is performed on both structural and functional data-based classification pipelines, as well as on the pipeline that is the fusion of both, in order to quantify the classification accuracy improvement in the presence of multimodal information. Furthemore, we performed statistical analysis with one-way analysis of variance (ANOVA) and post-hoc Tukey honest significant difference (HSD) tests in order to measure the results' statistical significance. Ultimately, we propose a qualitative analysis which compares our findings with the common ones in the clinical research. The method is validated on a multi-site, international Autism Brain Imaging Data Exchange I (ABIDE I) dataset, which consists of 1112 cases. We report a classification accuracy of 85.67% when using an ensemble of classifiers and analyze in detail the importance of a multimodal approach.

Keywords: Autism, ABIDE, deep learning, resting-state fMRI, structural MRI, classification

1. Introduction

According to the American Psychiatric Association, Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder characterized by persistent deficits in social communication (APA, 2013). It is said to be developmental since it tends to evolve in severity over time (Gotham et al., 2012; Szatmari et al., 2015). The symptoms generally appear in the first two years of life, and include, but are not restricted to, difficulty with communication and interaction, restricted interests and repetitive behaviors, as well as the degraded ability to function properly in various areas of life. The prevalence of ASD has been increasing over the past decades. In the United States, Xu et al. (2018) reported the prevalence of 0.67% in 2000, whereas APA (2013) estimated a prevalence of 1.47% in 2013. However, it is unclear whether higher rates reflect an expansion of the diagnostic criteria and increased awareness or a true increase in the frequency of ASD (APA, 2013). Furthermore, the average lifetime costs of ASD patient's treatment exceed one million dollars (Buescher et al., 2014).

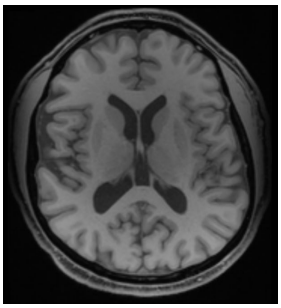
The cause of the disorder is still unknown to this date, although research suggests that it is most likely a result of a combination of factors that include genetics, brain structure and function, as well as environmental influences (APA, 2013). It is known that several risk factors increase the likelihood of having ASD, such as having a sibling with ASD, older parents, certain genetic conditions or a very low birth weight (Ha et al., 2015). Current diagnosis is interview-based, most commonly by conducting the Autism Diagnostic Observation Schedule (Lord et al., 1989) or the Autism Diagnostic Interview - Revised (Lord et al., 1994). Albeit being quite accurate, these methods are unable to point

out the biological basis behind behavioral symptoms, since the neuroanatomy is unclear (Subbaraju et al., 2017; Riddle et al., 2017). Furthermore, the diagnosis in adults is often more difficult, since symptoms can easily overlap with other mental health disorders (Ha et al., 2015). However, APA (2013) proposed some empirical differential diagnosis methods for separation of ASD from other overlapping neural disorders, such as Rett syndrome, selective mutism, language and social communication disorders, stereotypic movement disorder, attention-deficit/hyperactivity disorder and schizophrenia.

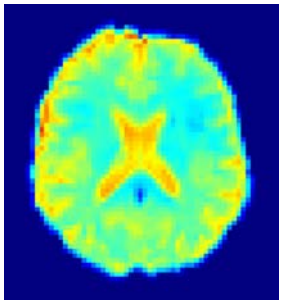
According to APA (2013), ASD consists of a variety of disorders that include (i) early infantile autism, (ii) childhood autism, (iii) Kanner's autism, (iv) high-functioning autism, (v) atypical autism, (vi) childhood disintegrative disorder, (vii) Asperger's Syndrome and (viii) pervasive developmental disorder not otherwise specified (PDD-NOS). However, last decades have seen an increase in works focusing on structural and functional brain abnormalities that would be symptomatic for the autism spectrum as a whole, and not on the specific part of the spectrum. Varying sized datasets have been used in order to observe common findings across the subjects with ASD in contrast to control groups. However, the findings typically do not hold over the whole set of ASD subjects, although MRI studies have provided many implications of neurodevelopmental characteristics underlying ASD (Ecker et al., 2015). Structural MRI studies usually focus on volumetric and morphometric analyses to examine abnormal brain anatomy, while functional MRI studies have tried to investigate connectivity patterns in the brain, both locally and globally.

This thesis presents a method that draws inspiration from previous research by Heinsfeld et al. (2018) and Kong et al. (2019), with the goal of improving the state-of-the-art classification results of ASD versus control group subjects. The method is evaluated using the large and international multi-site Autism Brain Imaging Data Exchange I (ABIDE I) dataset (Di Martino et al., 2014), which contains 1112 cases. Quantitative analysis of the results is conducted, for it allows the comparison with state-of-the-art results. Figure 1 shows an example of structural and averaged resting-state functional MRI series taken from the subject ID-51456 of the ABIDE I dataset. The principal hypothesis of the proposed method is that using multi-site data and combining both structural and functional information could potentially unveil patterns that have not been exploited so far, while at the same time improving generalization in terms of classification, due to the lack of reliance on a specific protocol.

The proposed method consists of several steps that include structural and functional data preprocessing, extraction of the features that are represented by connectivity matrices, the Fisher score as a feature dimensionality reduction technique and, ultimately, the classification of the data. Structural data connectivity matrices contain information about cortical gray matter volumes, as suggested by Kong et al. (2019), whereas functional data connectivity matrices contain information about correlation coefficients of mean blood-oxygen level dependent (BOLD) signals from pairs of regions of interest, as proposed by Heinsfeld et al. (2018). A BOLD signal is the magnetic resonance imaging contrast of blood deoxyhemoglobin activity, correlated with neural activity (Pelphrey, 2013).



(a) Axial slice extracted from structural MRI volume of an ASD subject (ID-51456)



(b) Corresponding slice taken from the averaged resting-state functional MRI volumes

Figure 1: Examples of structural and functional representation.

Statistical analysis with one-way ANOVA and post-hoc Tukey HSD tests was conducted on the obtained results in order to quantify the statistical significance. Additionally, we propose a qualitative analysis of the feature selection and ranking, with the goal of observing some connections between our method and clinical research findings.

2. State of the art

2.1. Structural MRI

Several studies tried to find the underlying ASD patterns in structural MRI. Voxel-based morphometry analysis (Riddle et al., 2017) showed an increase in total brain volume in children aged 2 to 4 with ASD, as well as an enlargement of the left anterior superior temporal gyrus. Other reports on total brain volume increase were based on measurements of the head circumference (Campbell et al., 1982; Hazlett et al., 2005) or accelerated postnatal brain growth (Piven et al., 1995; Courchesne et al., 2001). However, the picture is not so clear at a later age when it comes to volumetric analysis. While Aylward et al. (2002) observed no volumetric differences between ASD and control adult subjects, other studies concluded that the increase in total brain volume is still observable at a later age (Herbert et al., 2003; Palmen et al., 2005).

Other works investigated volumetric changes in particular regions of interest in brain, but also failed to

reach similar conclusions. Palmen et al. (2005) reported an increase in gray matter in all lobes of the brain, whereas Courchesne et al. (2007) observed an increase in gray matter volume particularly in temporal lobes. On the other side, Herbert et al. (2003) reported findings on increased white matter, while Palmen et al. (2005) noted no difference in ASD versus control subjects regarding white matter volume. Furthermore, Jou et al. (2011) reported a decrease in white matter volume in ASD patients. These inconsistent findings are most likely due to the small sample sizes or due to the fact that data was collected at a single site in each case (Riddle et al., 2017), since the acquisition site has significant effects on basic image properties (Nielsen et al., 2013; Castrillon et al., 2014). Reduced corpus callosum area is another common finding associated with ASD. However, recent studies have shown no difference in terms of the corpus callosum area when considering a multi-site dataset (Hiess et al., 2015). The same study confirmed a slight increase in total brain volume on average in subjects with ASD.

One promising study was based on the construction of an individual brain network for each subject in order to extract connectivity features between each pair of ROIs; those features were then ranked and used to perform ASD versus control classification via a DNN classifier (Kong et al., 2019). Features of interest were the cortical gray matter volumes in different regions of the cortex. Kong et al. (2019) reported an accuracy of 0.904, although the acquisition of the data was conducted on a single site. This is also, to our knowledge, the only notable study that tried to solve the classification problem based on structural information, whereas the vast majority of others tried to point out some of the common patterns among subjects with ASD versus the control group.

2.2. Resting-State Functional MRI

Resting-state functional MRI (rs-fMRI) was first described in 1995, when it was shown that low frequency oscillations in fMRI relate to spontaneous neural activity. The correlation of these low frequency fluctuations arises from fluctuations in blood oxygenation or flow. (Biswal et al., 1995). It is known that neurons do not contain intrinsic energy, and are instead provided with it by adjacent capillaries when activated, through a process called hemodynamic response (Lv et al., 2018). This results in a change of relative levels of oxyhemoglobin and deoxyhemoglobin that is consequently detected by fMRI imaging. However, the delay of the hemodynamic response following neural activation is responsible for the relatively poor temporal resolution of fMRI (Koch and Reid, 2012).

The lack of any task in rs-fMRI is particularly attractive for the investigation of brain disorders in patients that have difficulties performing certain task instructions. Functional connectivity in rs-fMRI is widely used to describe remote relationships in studies of the cerebral cortex parcellation and brain disorders (Jiang and Zuo, 2016). Further, rs-fMRI provides deeper insights in pathophysiology, which is what is lacking in the current diagnostic practice of ASD (Biswal et al., 1995; Kennedy and Courchesne, 2008). Thus, rs-fMRI can investigate, in a task-independent manner, the hypothesis that ASD involves disruptions of large-scale brain networks (Castelli et al., 2002; Belmonte et al., 2004).

Functional connectivity is expected to provide biomarkers for classifying brain disorders (Du et al., 2018). The principal idea is the usage of fMRI to detect brain networks among functionally interconnected regions. However, the main challenge when it comes to using functional connectivity for classification purposes is choosing the optimal strategy for feature selection. Connectivity matrices that contain functional correlations among different regions of the brain are very large, and get significantly larger in a voxel-wise connectivity analysis. If all those features are used, classifiers will tend to overfit. Furthermore, there is a problem of inevitable redundancy when having that many features. Moreover, the subjectivity of feature selection can also be an obstacle for result comparison (Heinsfeld et al., 2018). Another problem can be the separability of the classes when using functional connectivity matrices, which some research has tackled by projecting them in orthogonal directions using the Fukunaga-Koontz transform (Subbaraju et al., 2017).

Deep learning techniques have emerged as a recent trend (Plis et al., 2014; Iidaka, 2015; Calhoun and Sui, 2016; Ju et al., 2019). Popular approaches that yielded in good classification results include simple multi-layer perceptron (MLP) networks combined with the unsupervised training of stacked autoencoders (Kim et al., 2016; Guo et al., 2017). Another proposed approach involves graph convolutional networks, where nodes of the graph are image-based feature vectors, and edges represent phenotypic information (Parisot et al., 2018). However, the obtained accuracy of 70.4% does not surpass the ones obtained using the former, much more exploited method. The main drawbacks of deep learning techniques in general are computational costs, overfitting of the classifiers and interpretability of the results. However, with the rapid development of the technology, computational costs are becoming less of an issue, and overfitting can be somewhat overcome by applying regularization methods and feature dimensionality reduction. Furthermore, dimensionality reduction techniques can sometimes unravel some underlying patterns and tackle the interpretability problem up to a certain extent, by pointing out which neuroanatomical and neurofunctional alterations are of interest.

Since 2010, only 17 studies have used functional MRI data to perform the classification of ASD (Du et al., 2018). Figure 2 shows the classification accuracy ob-

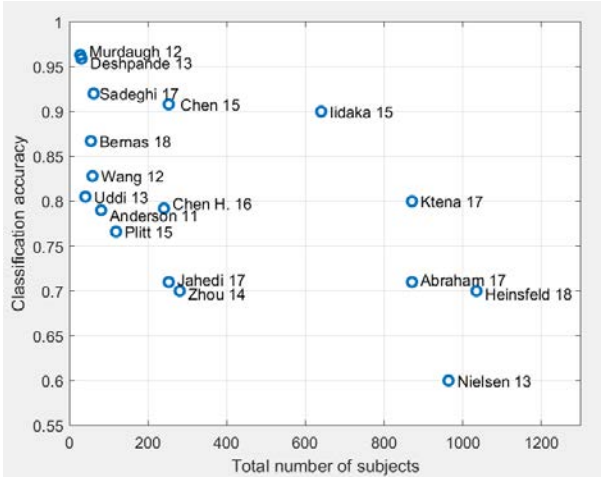


Figure 2: Overview and quantitative analysis of the state of the art (Anderson et al., 2011; Murdaugh et al., 2012; Wang et al., 2012; Deshpande et al., 2013; Nielsen et al., 2013; Uddin et al., 2013; Zhou et al., 2014; Chen et al., 2015; Iidaka, 2015; Plitt et al., 2015; Chen et al., 2016; Abraham et al., 2017; Jahedi et al., 2017; Ktena et al., 2018; Sadeghi et al., 2017; Bernas et al., 2018; Heinsfeld et al., 2018) on functional connectivity as a way of ASD classification. Figure modified from Du et al. (2018).

Table 1: Summary of number of subjects used from every screening site in each of the strategies conducted - functional, structural and combined classification pipelines. Rightmost column lists number of corresponding subjects in the original ABIDE I dataset, including the ones that failed the preprocessing step [ASD - Autism Spectrum Disorder, TC - Typical Control].

Site	Functional		Structural		Combined		ABIDE I	
	ASD	TC	ASD	TC	ASD	TC	ASD	TC
Caltech	19	18	17	18	17	17	19	19
CMU	3	2	14	12	3	2	14	13
KKI	12	27	20	32	11	26	22	33
Leuven	27	34	26	35	24	34	29	35
MAX MUN	18	24	21	31	17	22	24	33
NYU	73	98	74	103	69	96	79	105
OHSU	12	11	9	14	8	11	13	15
OLIN	14	11	17	16	11	11	20	16
PITT	22	23	27	24	20	20	30	27
SBL	14	12	14	14	13	11	15	15
SDSU	12	21	13	18	12	17	14	22
Stanford	17	19	19	16	16	15	20	20
Trinity	21	23	19	24	17	22	24	25
UCLA	36	39	50	43	34	37	62	47
UM	48	65	52	72	38	62	68	77
USM	38	23	56	42	37	23	58	43
Yale	22	26	27	25	21	23	28	28
Total	408	476	475	539	368	449	539	573

tained in each of the studies versus the sample size used. Even though several methods managed to obtain relatively high classification accuracy, there are some drawbacks of the proposed strategies that need to be addressed. Firstly, most of the studies used a small number of subjects to perform the classification. This tends to lead to unreliable results, because of the poor generalization. The real challenge is to replicate findings across large datasets. An accuracy above 0.9 is obtained only when using dozens of cases (Arbabshirani et al., 2017) and drops significantly when a larger dataset is introduced (Heinsfeld et al., 2018). Secondly, most of the studies used data acquired on a single site. This also does not generalize the problem efficiently, as the image properties highly depend on the imaging protocol conducted at each institution. Furthermore, only 4 of the mentioned studies used multi-site data with a number of subjects higher than 800, but only considered univariate approach; that is, only focused on functional findings, while neglecting structural information. Finally, only one study conducted a fusion approach, combining fMRI and diffusion tensor imaging (DTI) information, but on a sample size of only 30 subjects in total (Deshpande et al., 2013).

3. Material and methods

3.1. Dataset

The Autism Brain Imaging Data Exchange I (ABIDE I) dataset was used to conduct this study (Di Martino et al., 2014). It was released in August 2012 as a result of a collaboration involving 17 international sites and consists of 1112 cases, including 539 from individuals with ASD and 573 from typical controls, aged 7-64 with a median age of 14.7 years across the groups (Di Martino et al., 2014). The cases contain structural MRI images, resting-state fMRI series of images and a set of phenotypic information about subjects. Therefore, the dataset encompasses anatomical, functional and clinical data. Clinical data is, however, not used in this project, for some of it is missing from certain screening sites and is therefore not available in its entirety for the whole dataset. Table 1 summarizes the details of interest about the dataset, including separate listings for every acquisition site involved in the project.

Structural data and phenotypic information about subjects were obtained directly from the ABIDE I initiative. However, the preprocessed dataset that includes functional information was acquired from the Preprocessed Connectomes Project (PCP) (Craddock et al., 2013). All rs-fMRI series were subjected to processing pipeline called CPAC (Configurable Pipeline for the Analysis of Connectomes), which includes slice time correction, motion correction, intensity normalization, as well as band-pass filtering (0.01 Hz - 0.1 Hz) and spatial registration to the MNI152 template space. Various derivatives of the functional data are available at PCP, but the derivative of interest for the proposed classification pipeline is the time series of BOLD signals in different areas of the brain. Two different, commonly used atlases were tested - AAL (Automated Anatomical Labeling) atlas (Tzourio-Mazoyer et al., 2002) and CC200 (Cameron Craddock's 200 ROI) parcellation atlas (Craddock et al., 2012).

It is important to note, however, that the functional data is rendered useless in some cases, due to the motion artefacts, and is therefore not available from PCP.

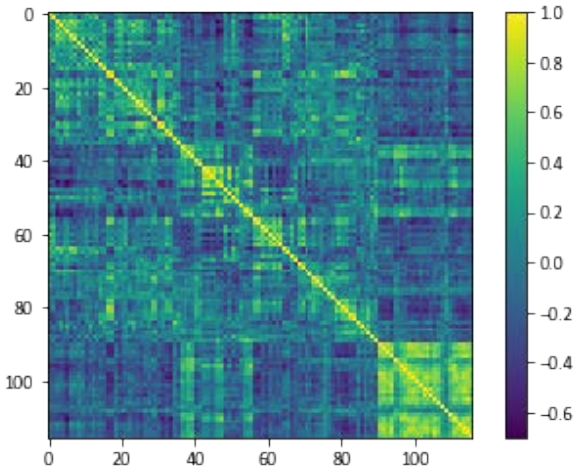
Motion artefacts are computed for each individual case using mean framewise displacement, and if it surpasses the value of 0.2, the corresponding subject is discarded. Mean framewise displacement is a measure of head motion, which compares the motion between current and previous volumes (Heinsfeld et al., 2018). This left us with a dataset of 884 rs-fMRI subjects, including 408 ASD patients and 476 control cases. As a summary, functional data for each subject is simply a set of time series of mean BOLD signal tracked in different regions of the brain, which are defined by AAL or CC200 atlas.

When it comes to structural information, the idea is somewhat similar. Cortical parcellation using Destrieux atlas (Destrieux et al., 2010) was performed for each MRI volume. To achieve this, the well-known Freesurfer software was used. The Freesurfer pipeline that was used to extract useful information involves multiple stages, most notable ones being motion correction, intensity normalization, skull stripping, registration of the volumes to a common space, segmentation and, ultimately, cortical parcellation. Apart from the division of cortex, a series of statistical measures, such as gray matter volume, cortical thickness or curvature, were computed for each of the parcels. However, due to the fact that serious motion artefacts are present in some of the structural MRI volumes, this processing was not possible. A total of 1014 cases were successfully processed, including 475 ASD patients and 539 control subjects. As a summary, structural data for each subject is a set of statistical measures for each of the cortex regions defined by Destrieux atlas.

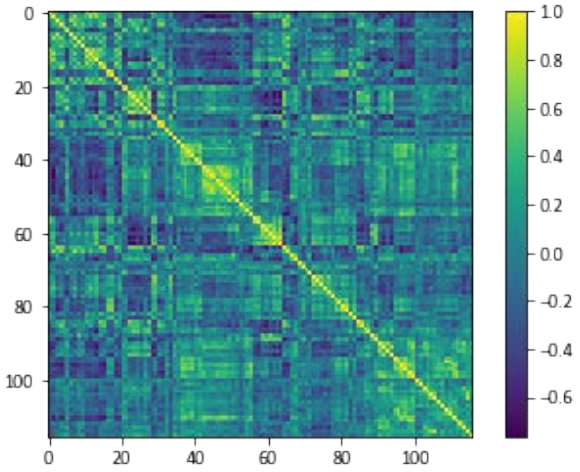
Finally, since the proposed classification strategy deals with structural and functional information, both separately and jointly (because we want to analyze the improvement of classification accuracy by incorporating data coming from different modalities), it is important to note that cross-referencing the remaining cases after preprocessing pipelines yields in 817 cases (368 ASD + 449 control) that are present in both subsets of the original dataset.

3.2. Functional data classification pipeline

Our pipeline for classification based on functional information was inspired by the approach described in Heinsfeld et al. (2018), as it was conducted on the whole ABIDE I dataset and showed promising results. Once the data was preprocessed with the CPAC pipeline and the time series of mean BOLD signals from different brain regions were extracted, the next step was to build a connectivity matrix. Such a matrix was constructed for each case individually and contained information about the correlation of BOLD series between each pair of the regions defined by an atlas (AAL atlas consists of 116 regions, whereas CC200 atlas consists of 200). Therefore, the dimensions of a connectivity matrix are 116-by-116 or 200-by-200, depending on the atlas used, and each element *ij* inside a matrix is the Pearson correlation



(a) Connectivity matrix of an ASD subject (ID-51456)



(b) Connectivity matrix of a control subject (ID-51476)

Figure 3: Examples of connectivity matrices of two subjects from Caltech subset of ABIDE I dataset constructed by using AAL atlas.

coefficient computed for the mean BOLD series from regions *i* and *j*. By the definition of Pearson correlation coefficient, elements of the matrix range from -1 to 1. All elements on the main diagonal are equal to 1, since they correspond to correlation of a signal with itself. Also, such a matrix is symmetrical, because of the commutative property of correlation coefficient computation. Examples of connectivity matrices are shown in Figure 3. We are, therefore, interested only in the upper triangle of the connectivity matrix, excluding the main diagonal, since the remainder of it is redundant. The part of interest is then flattened into a 1-dimensional vector for further manipulation. In case of the AAL atlas, such a vector contains 6670 elements, whereas in case of the CC200 atlas it contains 19900 elements.

Because of its high dimensionality, the feature vector is subjected to a dimensionality reduction technique, in order to get rid of highly correlated features, prevent overfitting and make the model more generalizable. A

technique used to achieve this goal is the Fisher score computation, which ranks the features in the order of distinctiveness and consequently decides which of them are of a lesser importance (Chen and Lin, 2006). It measures the discrimination of two sets of real numbers - the greater the score value, the higher the rank of a certain feature is. Given training vectors x_k, if the number of positive instances is n_+ and the number of negative instances is n_- (where positive and negative instances mean the ones belonging to one class or the other), Fisher score of the ith feature is:

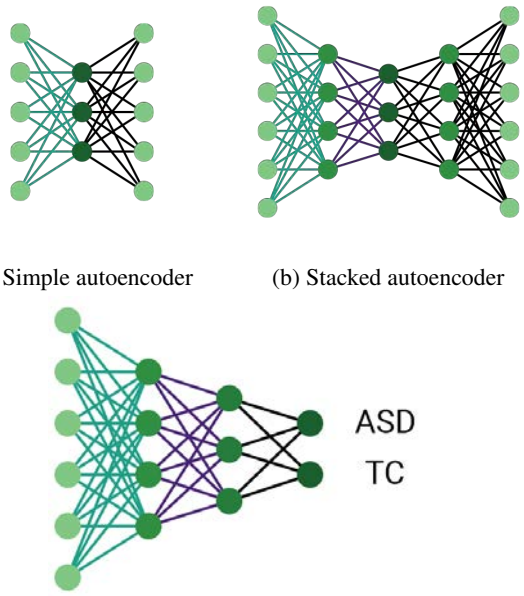
$$F(i) = \frac{(\bar{x}_i^+ - \bar{x}_i)^2 + (\bar{x}_i^- - \bar{x}_i)^2}{\frac{1}{n_+ - 1}\sum_{k=1}^{n_+}(x_{k,i}^+ - \bar{x}_i^+)^2 + \frac{1}{n_- - 1}\sum_{k=1}^{n_-}(x_{k,i}^- - \bar{x}_i^-)^2} \quad (1)$$

where $\bar{x}_i$, $\bar{x}_i^+$ and $\bar{x}_i^-$ are the mean of the ith feature of the whole, positive and negative sets respectively, $x_{k,i}^+$ is the ith feature of the kth positive instance, and $x_{k,i}^-$ is the ith feature of the kth negative instance (Kong et al., 2019). The numerator indicates inter-set discrimination, whereas the denominator indicates between-set discrimination.

What is left after application of the Fisher score is a reduced feature vector that serves as an input vector for the classifier. The classification step itself is done in two stages. First one consists of an unsupervised training of stacked autoencoders. An autoencoder is a simple network which tries to reconstruct the input as precisely as possible. Given the input vector, it tries to learn a lower-dimensional representation of it, from which it can then reconstruct the original vector. These two steps are referred to as encoding and decoding. Simply, it has an input layer, a hidden fully connected layer that encodes the input, and then a fully connected output layer that decodes the encoded representation. Parameters of the model are adjusted by back-propagation until the difference between input and output has been minimized.

A stacked autoencoder has a better learning ability and basically consists of two or more autoencoders. In the case of having two, the output of the first encoding stage is given as an input to the second autoencoder. Then, the decoding stage is done in a two-fold manner again - the second autoencoder decodes its input, and then the first autoencoder decodes the original input vector. An illustration of such a structure is shown in Figure 4.b.

The end result of the autoencoder training is apparent in the second stage of the classification step, which is a supervised training of a multilayer perceptron (MLP). The dataset is split into training, validation and testing sets and fed to a simple MLP with two hidden layers and the binary output layer. The number of nodes in the hidden layers corresponds to the number of nodes in the encoding layers of the stacked autoencoder. This ensures that the weights of the MLP can be initialized



(a) Simple autoencoder (b) Stacked autoencoder

(c) Multilayer perceptron

Figure 4: Graphical representation of the (a) simple and (b) stacked autoencoder structures and (c) multilayer perceptron. The colored weights of the encoding part of stacked autoencoder (b) are used as initializing weights of the MLP (c).

using the weights from the trained autoencoder, so that the MLP is able to learn hidden features from input vectors in the hidden layers, and then classify the subjects accordingly in the ultimate layer with softmax activation. This is illustrated in Figure 4.c, where the corresponding weights initialized from the autoencoder are color-coded in the same manner as in Figure 4.b. In order to prevent overfitting, dropout is introduced in the hidden layers, as well as additional regularization terms and batch normalization.

As a summary, the complete pipeline is illustrated in Figure 5.a, which shows all the notable steps, including the flattening of connectivity matrices into vectors, dimensionality reduction and training of the stacked autoencoder and the multi-layer perceptron.

3.3. Structural data classification pipeline

Our proposal for structural data-based classification was done in a similar manner. A connectivity matrix was built for each subject, and the upper triangular part was extracted and flattened to become a feature vector. The main difference between the functional and structural pipelines is the way connectivity matrices are built. Instead of computing the Pearson correlation coefficient, we are interested in relations between gray matter volumes in each pair of the cortical parcels defined by Destrieux atlas (148 regions, 74 in each hemisphere). Element ij of the matrix is the correlation between two parcels i and j which is defined by:

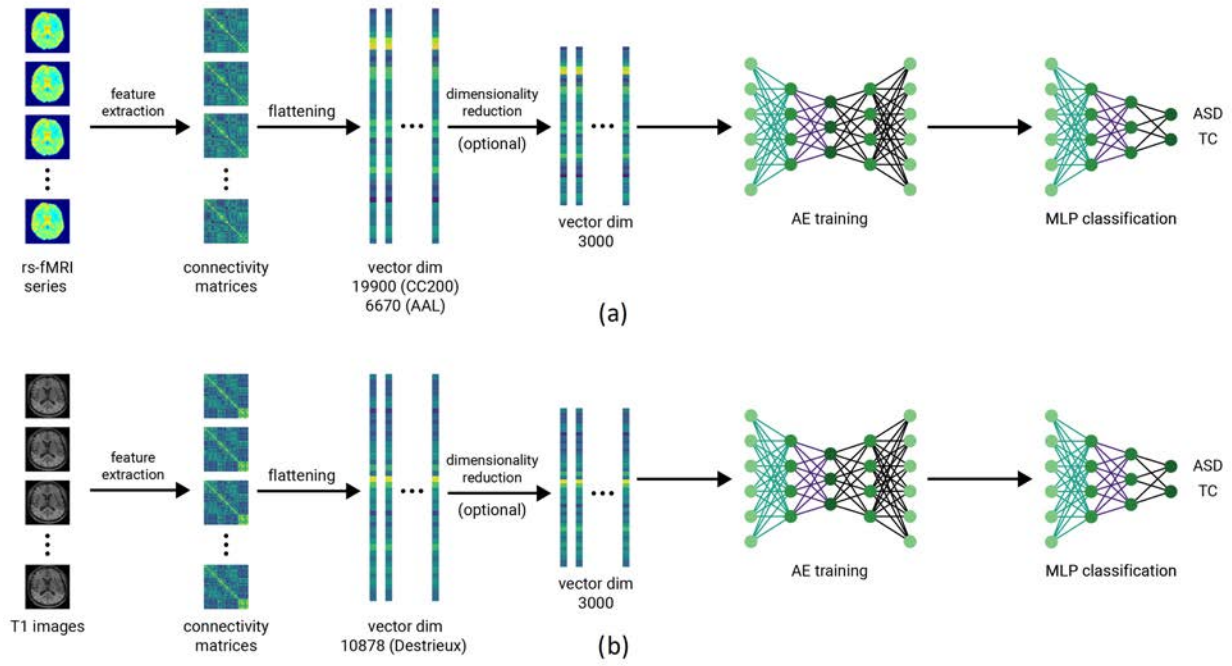


Figure 5: Graphical representation of the whole (a) functional and (b) structural separate data classification pipelines.

$$c(i, j) = \frac{1}{|gm(i) - gm(j)|^2 + 1} \tag{2}$$

where $gm(i)$ and $gm(j)$ are gray matter volumes of ROIs i and j (Kong et al., 2019).

As a result, flattened vector extracted from the connectivity matrix had 10878 elements prior to ranking them using Fisher score.

Fisher score was then used to reduce the dimensionality of the feature vectors, and the newly obtained ones were fed to the stacked autoencoders for unsupervised learning, and consequently to the MLP for the supervised learning and classification task. The pipeline is illustrated in Figure 5.b. This approach was inspired by the one described in Kong et al. (2019), and was taken for its similarity to the functional classification pipeline and the possibility to eventually merge the two together. Even though Kong et al. (2019) reported an accuracy of 90.39% on the ABIDE I dataset, the subset of cases used in the paper consisted of only 182 subjects, all taken from the NYU Langone Medical Center, making it a single-site study that is prone to worse generalization.

3.4. Combined data classification pipeline

One of the main contributions proposed in this master thesis involves combining the two previously described pipelines into one, with the goal of improving the classification results by accounting for different types of information. Since the functional and structural pipelines learn completely independent features, merging them together could possibly compensate errors to some extent. It is important to note that only cases which successfully underwent preprocessing pipelines, both functional and structural, can be considered as a part of the dataset for combined classification pipeline (in the ABIDE I dataset, 817 cases were successfully preprocessed). The merging was done using two different strategies.

First strategy simply involved concatenating structural and functional feature vectors after dimensionality reduction stages. Then, the classification was done using either newly obtained vector, or reducing its dimensionality again with the Fisher score to get another vector to be used as an input to the network. The classification stage itself remained unchanged, consisting of the unsupervised stacked autoencoder training, followed by the supervised training of an MLP. This approach is illustrated in Figure 6.

Second strategy consists of separate classification pipelines, as previously described and shown in Figure 5, followed by a decision-making method for choosing the final labels. In other words, we trained one autoencoder and multi-layer perceptron using the functional training and validation data, and the other set using the corresponding structural datasets. Then, test subjects were given as inputs to both classifiers, which led to having two separate labels. When they matched, the decision was to keep the obtained label; however, in case of a mismatch, the corresponding probabilities of softmax activations were compared, and the label with higher softmax probability was chosen.

The latter approach was explored further by, instead of having 2 classifiers, training an ensemble of a total of 10 classifiers (5 structural and 5 functional data based),

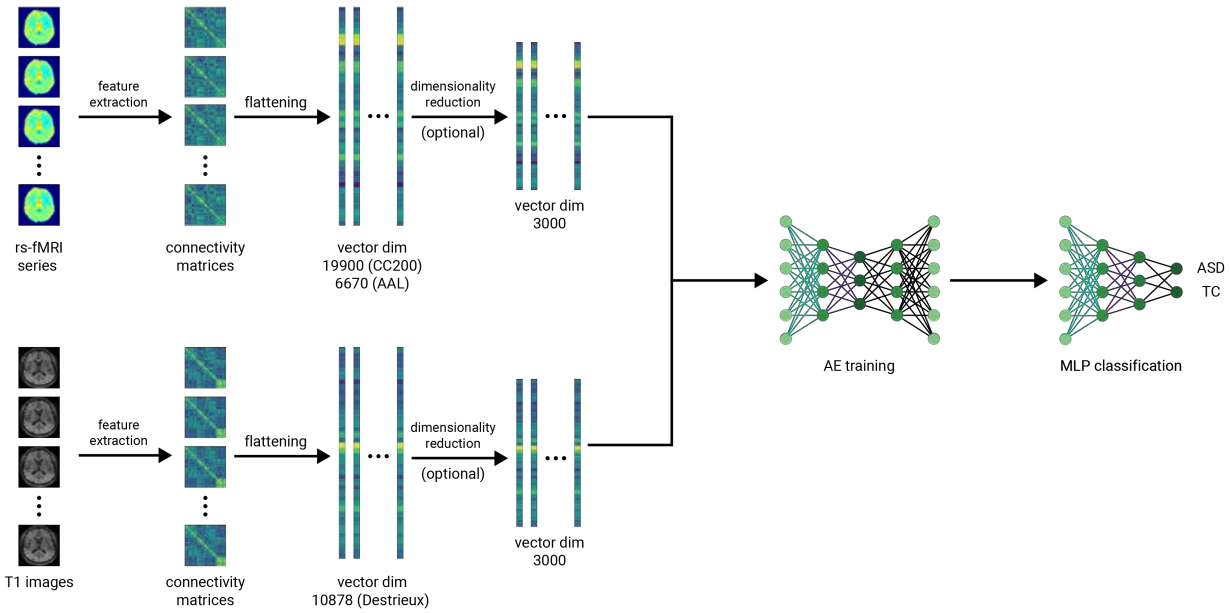


Figure 6: Graphical representation of the whole functional and structural combined data classification pipeline.

but changing the number of nodes in the hidden layers of autoencoders and MLPs, so that each one can learn different feature representations. In this case, each subject's feature vectors were given as input to all 10 classifiers, and the label assignment was conducted by either averaging 10 softmax activation probabilities or by majority voting. The underlying hypothesis is that additional classifiers with different ways of learning feature representations can add a certain margin of improvement in terms of classification accuracy and make the decision-making more robust (Kamnitsas et al., 2017).

3.5. Validation

Every model was validated by performing 10-fold cross-validation, similarly to the validation described in the papers by Heinsfeld et al. (2018) and Kong et al. (2019). In each fold, 10% of the corresponding dataset was used to test the classifier, while the remaining 90% of cases were used for training and validation, training encompassing 70% and validation 30% of the set. This allowed the evaluation of the model's robustness and behavioral effects when training and testing with different subsets of data. It is important to note that in each fold, the set was split in such a way that the subsets retained class balance and also contained cases coming from all or almost all screening sites; the idea being to keep the model as much generalizable as possible. Naturally, in case of training 2 classifiers or an ensemble of classifiers in the combined approach, the split into training, validation and testing sets was the same for all the classifiers in a certain fold.

The models were evaluated using the accuracy as a metric, which is the most common measure in the state of the art, including the research by Kong et al. (2019) and Heinsfeld et al. (2018) from which this thesis builds upon. This allowed for the quantitative comparison of our results with state of the art, although there is no guarantee that the dataset used in our method is identical to the ones the other researchers used. However, we tried to reproduce the results obtained in those two studies as a baseline, in order to quantify the results' potential improvement after merging the two modalities together.

Ultimately, we used the best model to analyze the accuracy obtained for each of the 17 screening sites individually. This allowed for some qualitative and quantitative comparison of the results with state of the art, especially since some of the previous research was conducted using only the data collected at one of the sites.

3.6. Statistical analysis

In order to perform the statistical analysis on the results obtained through the presented models, we performed one-way ANOVA (ANalysis Of VAriance) test, together with post-hoc Tukey HSD (Honestly Significant Difference) test, with the goal of indicating which models were significantly different from which. ANOVA and Tukey HSD tests were conducted separately on each of the three groups of models - functional data-based, structural data-based and combined data-based. 95% confidence interval was chosen, meaning that a p-value less than 0.05 indicates a high statistical significance of a certain result compared to another.

3.7. Qualitative analysis

In order to compare our method with common clinical findings, we propose a qualitative test of feature

ranking. Due to the fact that Fisher score ranks all the features by their distinctiveness, the idea is to observe whether the top ranked features correspond to discoveries noted in previous clinical and scientific research or not, and to check for the presence of potential common patterns.

According to Ha et al. (2015), the Default Mode Network (DMN), which is one of the commonly analyzed functional brain networks, shows a difference in brain activity between ASD and control subjects. DMN generally tends to be hypo-connected in adults with the disorder and hyper-connected in children with the same pathology. It comprises of several parts of the brain and includes the posterior cingulate gyrus, retrosplenial cortex, lateral parietal cortex, medial prefrontal cortex, superior frontal gyrus and temporal lobe. It has shown greater activity during resting-state functional MRI than during task-based screenings (Greicius et al., 2003), which is why it is of a particular interest for the ABIDE I dataset.

4. Results

4.1. Baseline implementation

Several experiments were conducted in order to verify the results' consistency with the ones obtained in the research done by Heinsfeld et al. (2018) and Kong et al. (2019). Functional data classification model was made using the entirety of ABIDE I dataset available at the Preprocessed Connectomes Project (884 cases, 408 ASD + 476 TC). We obtained a mean classification accuracy of 66.5%, whereas the one reported by Heinsfeld et al. (2018) was 70% (1035 cases, 505 ASD + 530 TC).

On the other hand, structural data classification pipeline was used on the ABIDE I subset acquired at the NYU Langone Medical Center, excluding the cases with serious motion artefacts that Freesurfer software was unable to process, which left us with 177 cases (74 ASD + 103 TC). The obtained accuracy of 83.04% was again lower than the one reported by Kong et al. (2019), which was 90.39% (182 cases, 78 ASD + 104 TC). The summary is shown in Table 2 (baseline implementation).

4.2. Functional data classification

In order to test the pipeline for functional data classification, four experiments were proposed. Two of those considered data preprocessed using the AAL atlas, whereas the other two used the data obtained with the CC200 atlas. In both approaches, we considered two sizes of input feature vectors - both without and with the application of dimensionality reduction method. When using the AAL atlas, the accuracies of 64.12% and 66.6% were reached without and with dimensionality reduction, respectively.

The accuracy was higher to a certain extent when using CC200 atlas. We report an accuracy of 66.5% without dimensionality reduction, and an accuracy of 71.27% when using Fisher score to reduce the size of the input vector.

The last model, that uses the CC200 atlas with the dimensionality reduction, was further explored by conducting the classification using an ensemble of 5 classifiers and averaging the softmax activations to obtain the final label. A slight improvement was reached, resulting in an accuracy of 71.95% (Table 2).

4.3. Structural data classification

When it comes to structural information, it was obtained using only the Destrieux atlas. Similarly to the previous approach, we tested two scenarios, both when using the original feature vector and the reduced one. The two obtained accuracies differ greatly, one being only 50.5% and the other, which included dimensionality reduction, being 76.13%.

Ensemble classification using 5 classifiers on the latter approach increased the accuracy to a greater extent than the one shown in the functional data approach. Averaging softmax activations in order to decide on the final label yielded in an average accuracy of 80.73% over the 10 folds of the cross-validation (Table 2).

4.4. Combined data classification

Combined approach was split into two major strategies, one considering the concatenation of the reduced vectors obtained via structural and functional data preprocessing, and the other considering separate classification of the functional and structural data, followed by the fusion of the obtained labels, either by averaging the softmax outputs or by majority voting.

Several models were proposed for both strategies, including the additional dimensionality reduction of the concatenated vector or classification using the ensemble of classifiers. It is important to note that the best result was obtained when considering the separate classification strategy, using an ensemble of 5 functional and 5 structural data classification models, followed by the averaging of all 10 softmax probabilities. We report an accuracy of 85.67% when conducting this approach.

Table 2 summarizes all of the conducted experiments and obtained accuracies, providing additional details on standard deviations over folds in cross-validation and peak accuracies (highest accuracies reached over 10 folds). For the best model, we performed a quantitative analysis of the accuracy reached for each of the 17 imaging sites. Table 3 summarizes the obtained values.

4.5. Statistical analysis

When it comes to functional data classification models (experiments 3, 4, 5 and 6 from Table 2), all of pairwise p-values were lower than 0.05. In terms of structural data classification models (experiments 7, 8 and 9

Table 2: Summary of the conducted experiments and the obtained results [Vector dim. - dimensionality of the input feature vectors, indicative of whether dimensionality reduction was used or not; Strategy - indication of whether the functional and structural feature vectors were concatenated prior to classification or classified separately; Ensemble - indication of whether an ensemble of classifiers was used or not, and if yes, how many classifiers were considered; * - label fusion was done using the average of softmax probabilities; ** - label fusion was done by majority voting, using average softmax probabilities only in case of a tie].

Baseline implementation							
Exp. no.	**Pipeline**	**Atlas**	**Cases**	**Vector dim.**	**Acc. mean [%]**	**Acc. std [%]**	**Peak acc. [%]**
1	functional	CC200	884	19900	66.50	5.23	76.40
2	structural	Destrieux	177	3000	83.04	4.31	89.47
Functional data classification							
3	functional	AAL	884	6670	64.12	4.09	70.79
4	functional	AAL	884	3000	66.60	5.23	74.16
5	functional	CC200	884	3000	71.27	4.44	77.01
6	ensemble of 5	CC200	884	3000	71.95	4.69	79.27
Structural data classification							
7	structural	Destrieux	1014	10878	50.50	3.63	54.90
8	structural	Destrieux	1014	3000	76.13	4.03	82.35
9	ensemble of 5	Destrieux	1014	3000	80.73	4.22	86.58
Combined data classification							
Exp. no.	**Strategy**	**Ensemble**	**Cases**	**Vector dim.**	**Acc. mean [%]**	**Acc. std [%]**	**Peak acc. [%]**
10	concatenate	no	817	6000	73.32	3.27	79.27
11	concatenate	no	817	3000	71.86	3.27	76.83
12	concatenate	yes (5)	817	6000	72.44	4.40	78.05
13	concatenate	yes (5)	817	3000	73.05	4.21	79.27
14	separate	no	817	3000	82.80	3.76	89.02
15	**separate***	**yes (5+5)**	**817**	**3000**	**85.67**	**3.80**	**91.46**
16	separate**	yes (5+5)	817	3000	85.42	2.94	90.24

Table 3: Classification accuracy obtained for each of the 17 screening sites using the best model, which considers a combined data classification approach with an ensemble of classifiers.

Site	No. of subjects	Accuracy [%]
Caltech	34	79.41
CMU	5	60.00
KKI	37	83.78
Leuven	58	89.66
MAX MUN	39	76.92
NYU	165	86.67
OHSU	19	78.95
OLIN	22	90.91
PITT	40	82.50
SBL	24	79.17
SDSU	29	86.21
Stanford	31	90.32
Trinity	39	84.62
UCLA	71	87.32
UM	100	88.00
USM	60	85.00
Yale	44	90.91

from Table 2), all 3 pairwise p-values were statistically significant ($p < 0.05$).

Lastly, in case of the combined approaches, we analyzed the statistical significance of the best model in comparison to the rest. The only reported p-value greater than 0.05 was obtained for the pair of experiments 15 and 16 from Table 2 and was equal to 0.8196; the rest were all lower than 0.05.

4.6. Qualitative functional feature analysis

Since Fisher score ranks the features according to how discriminant they are, we wanted to analyze whether or not the top features (i.e. top functional connectivity patterns between pairs of regions) correspond to common findings in clinical research based on functional connectivity. Table 4 lists the top 15 pairs of regions whose connectivity patterns are of the most interest when it comes to the classification task.

5. Discussion

The aim of this project was the improvement of ASD detection via means of combining structural and functional MRI information. Using our baseline implementations, which were based on the approaches developed

Table 4: Top 15 most discriminant pairs of regions for the classification task, corresponding to the top 15 functional connectivity features.

Feature rank	Pair of brain regions
1	Right Superior Frontal Gyrus Right Middle Temporal Gyrus
2	Left Superior Medial Gyrus Left Superior Frontal Gyrus
3	Right Middle Frontal Gyrus Right Insula Lobe
4	Right Inferior Frontal Gyrus Left Middle Occipital Gyrus
5	Right Caudate Nucleus Right Insula Lobe
6	Right Inferior Parietal Lobule Right Middle Frontal Gyrus
7	Right Anterior Cingulate Cortex Right Calcarine Gyrus
8	Left Inferior Frontal Gyrus Left Cerebellum
9	Right Fusiform Gyrus Left Insula Lobe
10	Right Middle Frontal Gyrus Left Subcallosal Gyrus
11	Right Inferior Parietal Lobule Left Cuneus
12	Right Thalamus Left Inferior Frontal Gyrus
13	Right Angular Gyrus Left Superior Frontal Gyrus
14	Left Thalamus Right Calcarine Gyrus
15	Left Precuneus Left Angular Gyrus

by Heinsfeld et al. (2018) and Kong et al. (2019) for functional and structural data classification respectively, we could not reach the exact same results as the ones presented in these two works. However, the results we obtained are still comparable and this discrepancy between our results and the reported ones can be justified through several aspects.

First of all, there is no guarantee that the cases we used for classification are the same ones used by Heinsfeld et al. (2018) and Kong et al. (2019). The only cases considered in this thesis are the ones that successfully underwent the functional or structural preprocessing pipelines. This means that the subjects with serious motion artefacts were discarded, which particularly reflected in the subset of the ABIDE I dataset coming from the CMU screening site; out of 27 cases, only 5 were available at PCP after functional preprocessing pipeline.

Another explanation for the results' inconsistency is the fact that the details of the implementation are not available in the two mentioned works. This means that some parameters of the network are left out, such as the number of nodes in hidden layers or the regularization strategies. Tuning these parameters can yield in different classification results. Finally, there is no certainty that the split of the dataset into training, validation and testing is the same, because of the fact that the cases are most likely not the same in the first place, but even if they were, the split is generated randomly.

It is observable that using the Fisher score as a dimensionality reduction technique helps improve the results, both when considering functional and structural pipelines separately, as well as in the combined approach. It prevents overfitting the classifier by selecting the most discriminant features from the defined set of features. Consequently, it minimizes redundancy. Dimensionality reduction is particularly beneficial when it comes to structural data classification. From Table 2, we can observe a jump from 50.5% to 76.13% classification accuracy just by removing the redundant features ($p < 0.05$). This difference is much lower when it comes to the functional data classification, but that pipeline also shows variation in results as a consequence of the atlas choice.

As shown, we used both AAL and CC200 atlases to preprocess the functional data. The CC200 atlas outperforms AAL in all of the conducted experiments (66.50% to 64.12% accuracy in experiments without dimensionality reduction; 71.27% to 66.60% accuracy in experiments that incorporated Fisher score), which is confirmed via statistical analysis and reported p-values less than 0.05. This may be due to the fact that CC200 atlas has 200 defined regions, whereas AAL has 116. More regions consequently unveil more information and connectivity patterns, which may have not been present or distinctive when using the AAL atlas.

If we consider separate approaches for structural and functional data classification, we can conclude that the structural pipeline outperforms the functional one in terms of the obtained accuracy. This can be due to the fact that there were more cases available for classification after the structural preprocessing pipeline. Furthermore, even though the features are defined independently in different pipelines, since they come from separate modalities, it is shown that the dimensionality reduction technique has a greater effect on the structural data classification than on the functional, which is another justification for the quantitative results' difference. In order words, the improvement in terms of classification accuracy is higher in the structural pipeline than in the functional when Fisher score is applied.

When it comes to implementation using the ensemble of classifiers, there is a statistically significant improvement in terms of classification accuracy in both functional and structural pipelines ($p < 0.05$). Each classifier in the ensemble is able to learn different representations of the input feature vectors, and by fusing the

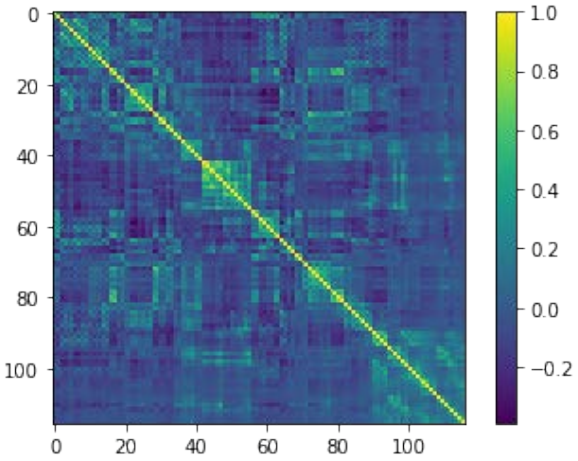
output labels together, errors can be compensated up to a certain extent. This compensation of errors is much more significant in case of the combined classification pipeline, because the input vectors are encompassing more information coming from different modalities. As far as the label fusion goes, we tried both majority voting and averaging the softmax activations. As it turns out, the latter approach slightly outperforms the former, because when conducting majority voting, all the labels are given the same weight, whereas when considering softmax probabilities, the classifiers that output probabilities with higher certainty are given more weight in the decision making than the ones with lower certainty. However, this improvement was not statistically significant, with the reported p-value of 0.8196.

Quantitative per site evaluation was performed to test the robustness of the best model and its generalization capability. The idea was to collect misclassified samples over 10 folds of the cross-validation and consequently find out which screening site they originated from. Then, the classification accuracies were computed for each of the 17 sites. The lowest accuracy of 60% was obtained for the CMU site, but the total number of cases considered from that particular site is only 5, which is significantly lower than the number of cases coming from other sites. Having more available cases would arguably increase this site's accuracy in individual site analysis. In case of having a low number of cases, 1 or 2 errors have a greater impact on classification accuracy than the same number of errors in case of having a larger dataset.

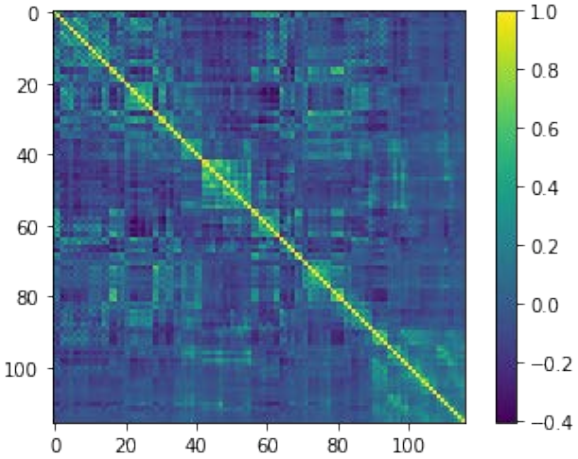
In order to assess our methodology in terms of the common findings in clinical research, we performed two qualitative analyses. First one was a comparison of the average functional connectivity matrices coming from ASD subjects versus control group subjects. The idea was to see if there are observable functional patterns between the two classes, and if so, do they correspond to the previous works' findings. Figure 7 shows the two mean matrices. Qualitatively, there is little to no difference between the two. This could mean that having a lot of different cases coming from several screening sites with varying protocols can, on average, neutralize the functional connectivity differences. Also, another hypothesis is that there are underlying groups of patterns not observable simply by evaluating correlations between pairs of regions of interest. However, the model is able to extract those features when it comes to binary classification, which shows in the obtained quantitative results.

Another analysis we performed tried to investigate whether or not the top discriminant features ranked by Fisher score correspond to the common findings in functional brain connectivity, particularly in the research related to Default Mode Network connectivity.

By comparing the DMN regions with the regions corresponding to the top ranked features (shown in Table



(a) ASD



(b) TC

Figure 7: Average functional connectivity matrices of the two classes obtained using the AAL atlas.

4), some similarities can be observed. There is a presence of superior frontal gyri and temporal lobe in both, but there are also pairs of regions that the classification network is able to discriminate well, that are not present in DMN. The reasoning is that functional connectivity patterns are observed globally, not by focusing on one particular brain network. Our method quantifies the connectivity of all the ROI pairs, and consequently selects the most distinct connections in order to optimize classification accuracy, whereas the DMN analyses focus only on connectivity of the regions that fall under the definition of DMN, disregarding the rest. This selection of the most discriminant features can also justify the improvement in the accuracy when conducting the combined approach; we consider two different modalities, with two separate types of features, selecting the most distinguishable ones from both.

We conducted an analogous analysis to examine top features in case of structural data. However, the way features are extracted and presented does not allow for

any comparison with other work nor for drawing any parallels with common clinical findings, since the features we used were described in work by Kong et al. (2019) and, to our knowledge, were only used in that research for these particular purposes. Nevertheless, there are some observations and recurring patterns that can be noted. For instance, in the top 15 features (i.e. top 15 most discriminant correlations between pairs of cortical regions), left transverse temporal sulcus appears in 5. This can signify some importance of that particular cortical parcel when it comes to ASD classification, since it is 1 of possible 148 parcels in total that appears in 5 of the top 15 features. Other recurring regions on the list are left intraparietal sulcus and transverse parietal sulci that appear 4 times, as well as right subcallosal gyrus, which appears in 3 features.

It is also important to note that, as far as the implementation choices go, the number of nodes in the hidden layers of the default autoencoder and MLP (which is 1000 and 600 in two layers) was used in the work of Heinsfeld et al. (2018) and was kept in our approach. The same architecture was then preserved for the structural data classification pipeline. On the other hand, the choice to lower the input vectors' dimensionality to 3000 was shown to be the best one in the paper by Kong et al. (2019) (different models were tested by varying dimensionality of the input vectors from 2000 to 5000); consequently, we applied the same reduction constraint when conducting the functional data classification strategy.

295 cases out of 1112 in the original dataset did not meet the required preprocessing criteria, either on functional or structural part, or in some cases both. This is potentially one of the main drawbacks, because the reported results are not obtained using the full dataset. Furthermore, this restriction obstructs the comparison with some of the other works on the same topic, because the subset we used does not necessarily correspond to the ones used in other papers, which particularly reflected in our baseline implementation, where we tried to reproduce results obtained in the works by Heinsfeld et al. (2018) and Kong et al. (2019). The results we reported were somewhat lower, most likely due to differences in datasets, parameters of the classification networks or the split into training, validation and testing sets. Another drawback of the proposed method is the fact that preprocessing step is time and resource consuming, especially when it comes to structural data segmentation and computation of statistics for the obtained cortical parcels. The Freesurfer pipeline was taking approximately 20 to 24 hours to process a single case, while processing 42 batches of subjects in parallel on a 60-core processor. Considering that the total number of subjects in the ABIDE I dataset is 1112, the structural data preprocessing took around 1 month to complete.

However, we were able to obtain arguably high classification results in comparison to the other works, even though the data originated from 17 different screening sites and was acquired using different protocols. This means that our method has a good generalization ability and does not rely on a specific protocol.

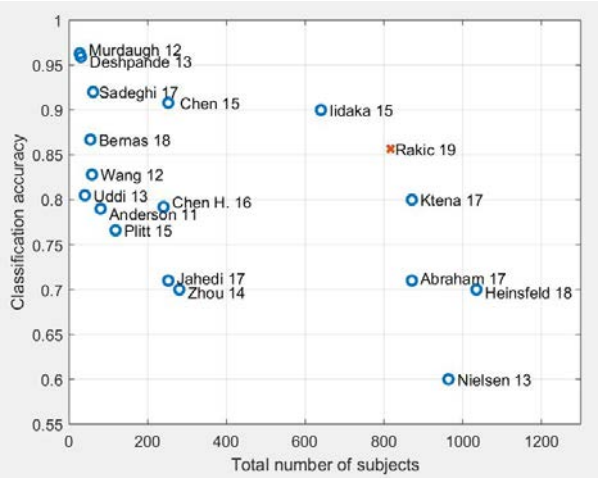


Figure 8: Comparison of the results obtained using the combined functional and structural data classification pipeline with the ones reached in previous works.

Additionally, current diagnosis of ASD is based on 2 main criteria: impairments in social communication and interaction and a restrictive, repetitive range of interests, behaviors and activities (APA, 2013). An unexperienced clinician is likely to incorrectly apply the criteria for autism and related conditions, which is a major concern. Another significant problem in current clinical practice is the delayed diagnosis, since early initialization of treatment increases probability for a favorable outcome.

Taking this into consideration, our method may provide additional insight when it comes to ASD diagnosis. Even though the neuroimaging studies in the field yielded in inconsistent results and are still not considered a diagnostic tool, our method may be used as a stand-alone tool for ASD detection, as it may potentially unveil some useful patterns and findings for discrimination of the disorder.

Finally, Figure 8 shows how our best model compares to the already presented overview of the previous works that dealt with classification of ASD. Even though the accuracy we obtained is not the highest one, every other result that is quantitatively better was based on a smaller dataset, often including data originating from only a single screening site. Furthermore, the works based on a larger dataset than the one we used yielded in a significantly lower accuracy. However, the behavior of our model when tested on the entirety of ABIDE I dataset remains unknown, since a portion of subjects was discarded due to various artefacts.

6. Conclusions

In this master thesis we proposed a method for classification of Autism Spectrum Disorder versus control group. The proposed method, based on a network consisted of autoencoders and multi-layer perceptrons, was tested on both functional and structural data (in both separate and combined manner) available from ABIDE I dataset. Our proposal was inspired by the works of Heinsfeld et al. (2018) and Kong et al. (2019), which dealt with the classification of ASD based on functional and structural data, respectively. The classification task itself was done in a similar manner in both of those papers, even though the approach of Heinsfeld et al. (2018) did not include any dimensionality reduction technique, and the approach of Kong et al. (2019) was based only on ABIDE I subset coming from NYU Langone Medical Center screening site. This opened up the possibility to incorporate both structural and functional information and analyze the potential improvement in terms of classification accuracy.

We showcased the importance of the multimodal approach by analyzing the obtained results qualitatively and quantitatively. By encompassing different types of information in our classification algorithm, we were able to improve the results in a statistically significant manner, which was shown through the analysis with one-way ANOVA and post-hoc Tukey HSD tests. The highest obtained classification accuracy of 85.67% was a result of a multimodal strategy that included en ensemble of classifiers for both structural and functional data classification. This model that yielded in best results in terms of accuracy was further explored by performing per-site analysis.

Furthermore, we analyzed the impact of feature dimensionality reduction technique in a two-fold manner. From one perspective, it served to prevent overfitting of the classifier and to control redundancy of the features; from another, ranking the features and selecting the top most discriminant ones allowed for interpretability of the features to a certain extent. For instance, we were able to observe some common findings between our top features from functional dataset and a commonly analyzed Default Mode Network in resting-state functional MRI research. On the other hand, we noted some recurring patterns in top features of structural data, which could suggest the importance of certain cortical parcels, such as transverse temporal sulcus, in ASD classification task.

As a summary, even though some implementation differences do not allow a direct quantitative comparison of our results with the ones obtained in other state-of-the-art works, the proposed approach shows that incorporation of different modalities and types of information significantly improves classification accuracy of ASD versus typical controls.

7. Acknowledgments

We would like to thank Kaisar Kushibar for providing technical support and help with the usage of the Freesurfer software. Additionally, special thanks goes to Marcio Rockenbach for his feedback and constructive criticism of the thesis. Ultimately, we thank Petar Brković for the design support and coming up with the graphical solutions for the thesis.

References

Abraham, A., Milham, M.P., Di Martino, A., Craddock, R.C., Samaras, D., Thirion, B., Varoquaux, G., 2017. Deriving reproducible biomarkers from multi-site resting-state data: an autism-based example. NeuroImage 147, 736–745.

Anderson, J.S., Nielsen, J.A., Froehlich, A.L., DuBray, M.B., Druzgal, T.J., Cariello, A.N., Cooperrider, J.R., Zielinski, B.A., Ravichandran, C., Fletcher, P.T., et al., 2011. Functional connectivity magnetic resonance imaging classification of autism. Brain 134, 3742–3754.

APA, 2013. Diagnostic and statistical manual of mental disorders (DSM-5®). American Psychiatric Pub.

Arbabshirani, M.R., Plis, S., Sui, J., Calhoun, V.D., 2017. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. NeuroImage 145, 137–165.

Aylward, E.H., Minshew, N.J., Field, K., Sparks, B., Singh, N., 2002. Effects of age on brain volume and head circumference in autism. Neurology 59, 175–183.

Belmonte, M.K., Allen, G., Beckel-Mitchener, A., Boulanger, L.M., Carper, R.A., Webb, S.J., 2004. Autism and abnormal development of brain connectivity. Journal of Neuroscience 24, 9228–9231.

Bernas, A., Aldenkamp, A.P., Zinger, S., 2018. Wavelet coherence-based classifier: a resting-state functional mri study on neurodynamics in adolescents with high-functioning autism. Computer Methods and Programs in Biomedicine 154, 143–151.

Biswal, B., Zerrin Yetkin, F., Haughton, V.M., Hyde, J.S., 1995. Functional connectivity in the motor cortex of resting human brain using echo-planar mri. Magnetic Resonance in Medicine 34, 537–541.

Buescher, A.V., Cidav, Z., Knapp, M., Mandell, D.S., 2014. Costs of autism spectrum disorders in the united kingdom and the united states. The Journal of the American Medical Association Pediatrics 168, 721–728.

Calhoun, V.D., Sui, J., 2016. Multimodal fusion of brain imaging data: a key to finding the missing link (s) in complex mental illness. Biological Psychiatry: Cognitive Neuroscience and Neuroimaging 1, 230–244.

Campbell, M., Rosenbloom, S., Perry, R., George, A., Kricheff, I., Anderson, L., Small, A., Jennings, S., 1982. Computerized axial tomography in young autistic children. The American Journal of Psychiatry .

Castelli, F., Frith, C., Happé, F., Frith, U., 2002. Autism, asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. Brain 125, 1839–1849.

Castrillon, J.G., Ahmadi, A., Navab, N., Richiardi, J., 2014. Learning with multi-site fmri graph data, in: 2014 48th Asilomar Conference on Signals, Systems and Computers, IEEE. pp. 608–612.

Chen, C.P., Keown, C.L., Jahedi, A., Nair, A., Pflieger, M.E., Bailey, B.A., Müller, R.A., 2015. Diagnostic classification of intrinsic functional connectivity highlights somatosensory, default mode, and visual regions in autism. NeuroImage: Clinical 8, 238–245.

Chen, H., Duan, X., Liu, F., Lu, F., Ma, X., Zhang, Y., Uddin, L.Q., Chen, H., 2016. Multivariate classification of autism spectrum disorder using frequency-specific resting-state functional connectivitya multi-center study. Progress in Neuro-Psychopharmacology and Biological Psychiatry 64, 1–9.

Chen, Y.W., Lin, C.J., 2006. Combining svms with various feature selection strategies, in: Feature extraction. Springer, pp. 315–324.

Courchesne, E., Karns, C., Davis, H., Ziccardi, R., Carper, R., Tigue, Z., Chisum, H., Moses, P., Pierce, K., Lord, C., et al., 2001. Unusual brain growth patterns in early life in patients with autistic disorder: an mri study. Neurology 57, 245–254.

Courchesne, E., Pierce, K., Schumann, C.M., Redcay, E., Buckwalter, J.A., Kennedy, D.P., Morgan, J., 2007. Mapping early brain development in autism. Neuron 56, 399–413.

Craddock, C., Benhajali, Y., Chu, C., Chouinard, F., Evans, A., Jakab, A., Khundrakpam, B.S., Lewis, J.D., Li, Q., Milham, M., et al., 2013. The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. Neuroinformatics 4.

Craddock, R.C., James, G.A., Holtzheimer III, P.E., Hu, X.P., Mayberg, H.S., 2012. A whole brain fmri atlas generated via spatially constrained spectral clustering. Human Brain Mapping 33, 1914–1928.

Deshpande, G., Libero, L., Sreenivasan, K.R., Deshpande, H., Kana, R.K., 2013. Identification of neural connectivity signatures of autism using machine learning. Frontiers in Human Neuroscience 7, 670.

Destrieux, C., Fischl, B., Dale, A., Halgren, E., 2010. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. Neuroimage 53, 1–15.

Di Martino, A., Yan, C.G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M., et al., 2014. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular Psychiatry 19, 659.

Du, Y., Fu, Z., Calhoun, V.D., 2018. Classification and prediction of brain disorders using functional connectivity: Promising but challenging. Frontiers in Neuroscience 12.

Ecker, C., Bookheimer, S.Y., Murphy, D.G., 2015. Neuroimaging in autism spectrum disorder: brain structure and function across the lifespan. The Lancet Neurology 14, 1121–1134.

Gotham, K., Pickles, A., Lord, C., 2012. Trajectories of autism severity in children using standardized ados scores. Pediatrics 130, e1278–e1284.

Greicius, M.D., Krasnow, B., Reiss, A.L., Menon, V., 2003. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. Proceedings of the National Academy of Sciences 100, 253–258.

Guo, X., Dominick, K.C., Minai, A.A., Li, H., Erickson, C.A., Lu, L.J., 2017. Diagnosing autism spectrum disorder from brain resting-state functional connectivity patterns using a deep neural network with a novel feature selection method. Frontiers in Neuroscience 11, 460.

Ha, S., Sohn, I.J., Kim, N., Sim, H.J., Cheon, K.A., 2015. Characteristics of brains in autism spectrum disorder: structure, function and connectivity across the lifespan. Experimental Neurobiology 24, 273–284.

Hazlett, H.C., Poe, M., Gerig, G., Smith, R.G., Provenzale, J., Ross, A., Gilmore, J., Piven, J., 2005. Magnetic resonance imaging and head circumference study of brain size in autism: birth through age 2 years. Archives of General Psychiatry 62, 1366–1376.

Heinsfeld, A.S., Franco, A.R., Craddock, R.C., Buchweitz, A., Meneguzzi, F., 2018. Identification of autism spectrum disorder using deep learning and the abide dataset. NeuroImage: Clinical 17, 16–23.

Herbert, M., Ziegler, D., Deutsch, C., Obrien, L., Lange, N., Bakardjiev, A., Hodgson, J., Adrien, K., Steele, S., Makris, N., et al., 2003. Dissociations of cerebral cortex, subcortical and cerebral white matter volumes in autistic boys. Brain 126, 1182–1192.

Hiess, R.K., Alter, R., Sojoudi, S., Ardekani, B., Kuzniecky, R., Pardoe, H., 2015. Corpus callosum area and brain volume in autism spectrum disorder: quantitative analysis of structural mri from the abide database. Journal of Autism and Developmental Disorders 45, 3107–3114.

Iidaka, T., 2015. Resting state functional magnetic resonance imaging and neural network classified autism and control. Cortex 63, 55–67.

Jahedi, A., Nasamran, C.A., Faires, B., Fan, J., Müller, R.A., 2017. Distributed intrinsic functional connectivity patterns predict diagnostic status in large autism cohort. Brain Connectivity 7, 515–525.

Jiang, L., Zuo, X.N., 2016. Regional homogeneity: a multimodal, multiscale neuroimaging marker of the human connectome. The Neuroscientist 22, 486–505.

Jou, R.J., Mateljevic, N., Minshew, N.J., Keshavan, M.S., Hardan, A.Y., 2011. Reduced central white matter volume in autism: Implications for long-range connectivity. Psychiatry and Clinical Neurosciences 65, 98–101.

Ju, R., Hu, C., Zhou, P., Li, Q., 2019. Early diagnosis of alzheimer's disease based on resting-state brain networks and deep learning. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 16, 244–257.

Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., Rajchl, M., Lee, M., Kainz, B., Rueckert, D., et al., 2017. Ensembles of multiple models and architectures for robust brain tumour segmentation, in: International MICCAI Brainlesion Workshop, Springer. pp. 450–462.

Kennedy, D.P., Courchesne, E., 2008. The intrinsic functional organization of the brain is altered in autism. Neuroimage 39, 1877–1885.

Kim, J., Calhoun, V.D., Shim, E., Lee, J.H., 2016. Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: Evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. Neuroimage 124, 127–146.

Koch, C., Reid, R.C., 2012. Neuroscience: Observatories of the mind. Nature 483, 397.

Kong, Y., Gao, J., Xu, Y., Pan, Y., Wang, J., Liu, J., 2019. Classification of autism spectrum disorder by combining brain connectivity and deep neural network classifier. Neurocomputing 324, 63–68.

Ktena, S.I., Parisot, S., Ferrante, E., Rajchl, M., Lee, M., Glocker, B., Rueckert, D., 2018. Metric learning with spectral graph convolutions on brain connectivity networks. NeuroImage 169, 431–442.

Lord, C., Rutter, M., Goode, S., Heemsbergen, J., Jordan, H., Mawhood, L., Schopler, E., 1989. Austism diagnostic observation schedule: A standardized observation of communicative and social behavior. Journal of Autism and Developmental Disorders 19, 185–212.

Lord, C., Rutter, M., Le Couteur, A., 1994. Autism diagnostic interview-revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. Journal of Autism and Developmental Disorders 24, 659–685.

Lv, H., Wang, Z., Tong, E., Williams, L., Zaharchuk, G., Zeineh, M., Goldstein-Piekarski, A., Ball, T., Liao, C., Wintermark, M., 2018. Resting-state functional mri: everything that nonexperts have always wanted to know. American Journal of Neuroradiology 39, 1390–1399.

Murdaugh, D.L., Shinkareva, S.V., Deshpande, H.R., Wang, J., Pennick, M.R., Kana, R.K., 2012. Differential deactivation during mentalizing and classification of autism based on default mode network connectivity. Public Library of Science One 7, e50064.

Nielsen, J.A., Zielinski, B.A., Fletcher, P.T., Alexander, A.L., Lange, N., Bigler, E.D., Lainhart, J.E., Anderson, J.S., 2013. Multisite functional connectivity mri classification of autism: Abide results. Frontiers in Human Neuroscience 7, 599.

Palmen, S.J., Pol, H.E.H., Kemner, C., Schnack, H.G., Durston, S., Lahuis, B.E., Kahn, R.S., Van Engeland, H., 2005. Increased gray-matter volume in medication-naive high-functioning children with autism spectrum disorder. Psychological Medicine 35, 561–570.

Parisot, S., Ktena, S.I., Ferrante, E., Lee, M., Guerrero, R., Glocker, B., Rueckert, D., 2018. Disease prediction using graph convolutional networks: Application to autism spectrum disorder and alzheimers disease. Medical Image Analysis 48, 117–130.

Pelphrey, K.A., 2013. Blood-oxygen-level-dependent (bold) signal. Encyclopedia of Autism Spectrum Disorders , 465–466.

Piven, J., Arndt, S., Bailey, J., Havercamp, S., et al., 1995. An mri study of brain size in autism. The American Journal of Psychiatry 152, 1145.

Plis, S.M., Hjelm, D.R., Salakhutdinov, R., Allen, E.A., Bockholt, H.J., Long, J.D., Johnson, H.J., Paulsen, J.S., Turner, J.A., Calhoun, V.D., 2014. Deep learning for neuroimaging: a validation study. Frontiers in Neuroscience 8, 229.

Plitt, M., Barnes, K.A., Martin, A., 2015. Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards. NeuroImage: Clinical 7, 359–366.

Riddle, K., Cascio, C.J., Woodward, N.D., 2017. Brain structure in autism: a voxel-based morphometry analysis of the autism brain imaging database exchange (abide). Brain Imaging and Behavior 11, 541–551.

Sadeghi, M., Khosrowabadi, R., Bakouie, F., Mahdavi, H., Eslahchi, C., Pouretemad, H., 2017. Screening of autism based on task-free fmri using graph theoretical approach. Psychiatry Research: Neuroimaging 263, 48–56.

Subbaraju, V., Suresh, M.B., Sundaram, S., Narasimhan, S., 2017. Identifying differences in brain activities and an accurate detection of autism spectrum disorder using resting state functional-magnetic resonance imaging: A spatial filtering approach. Medical Image Analysis 35, 375–389.

Szatmari, P., Georgiades, S., Duku, E., Bennett, T.A., Bryson, S., Fombonne, E., Mirenda, P., Roberts, W., Smith, I.M., Vaillancourt, T., et al., 2015. Developmental trajectories of symptom severity and adaptive functioning in an inception cohort of preschool children with autism spectrum disorder. The Journal of the American Medical Association Psychiatry 72, 276–283.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. Neuroimage 15, 273–289.

Uddin, L.Q., Supekar, K., Lynch, C.J., Khouzam, A., Phillips, J., Feinstein, C., Ryali, S., Menon, V., 2013. Salience network–based classification and prediction of symptom severity in children with autism. The Journal of the American Medical Association Psychiatry 70, 869–879.

Wang, H., Chen, C., Fushing, H., 2012. Extracting multiscale pattern information of fmri based functional brain connectivity with application on classification of autism spectrum disorders. Public Library of Science One 7, e45502.

Xu, G., Strathearn, L., Liu, B., Bao, W., 2018. Prevalence of autism spectrum disorder among us children and adolescents, 2014-2016. The Journal of the American Medical Association 319, 81–82.

Zhou, Y., Yu, F., Duong, T., 2014. Multiparametric mri characterization and prediction in autism spectrum disorder using graph theory and machine learning. Public Library of Science One 9, e90405.

Medical Imaging and Applications

Master Thesis, June 2019

DeepDraw! Developing a web application for medical image annotation and computer aided analysis

Zafar Toshpulatov, Robert Marti, Oliver Diaz

Computer Vision and Robotics Research Group of the University of Girona, Catalonia, Spain

Abstract

The analyzing and collection of large amount medical data with labeled groundtruth is the biggest challenge at present. To guarantee accurate and valid groundtruth, medical experts are required to annotate medical image datasets. Recently, Deep Learning has become widely popular in medical image analysis. In this paper, we present computer-aided medical image annotation tool called DeepDraw, based on web application. This interactive web application integrates two tools: manual annotation that performs the annotation by drawing manually by the user; intelligent annotation tool that accomplishes fully automatic segmentation of provided medical image using well-known U-Net deep learning architecture. In addition, after applying intelligent annotation tool, annotated contour can also be corrected manually by the medical expert. We adopted the OHIF web based medical image viewer platform to our work and the Orthanc mini PACS system was connected to the platform. One of the conveniences of the platform is to register the users. Moreover, the REST backend APIs were built to run the deep learning model and store the annotation into the database. In order to do this task, the INbreast digital mammography dataset was used which contains of 107 images of mass lesions. The results shows that a web platform allows medical experts to annotate huge and complex medical image collections much faster.

Keywords: Medical image annotation, Web application, Digital mammography, Deep Learning, Segmentation, U-Net

1. Introduction

Internet technologies have evolved gradually to such an extent that it has become possible to build web applications comparable with commonly used desktop applications. Web applications are used in many different areas, including education, media, business and the medical community among others. Recently, many radiologists have been involved with data scientists in the development of these web applications for radiological purposes. By developing web-based tools, users can access medical imaging processing platform wherever the Internet is exist. Although a large number of medical web applications, there are still some technological challenges that need to be solved such as medical annotation applications for labeling lesions or organs in human body. A growing number of medical image data requires computer-assisted medical image annotation (Qiusha Min and Liu, 2018).

The challenges of medical image analysis and computer-assisted intervention are being addressed in recent years with the aid of machine learning approaches. Currently, researchers are mainly using machine learning techniques to develop tools. Machine learning is a set of algorithmic methods that allow computer systems to make data-driven predictions from big data. These methods have many applications that can be adapted to the field of medicine (Akkus et al., 2017). The developed machine learning platforms are flexible, but they don't provide certain functions for analyzing medical images, and adapting them for this application requires considerable implementation efforts. Most of the proposed machine learning methods so far do not use explicit dependencies between annotations (Eli Gibson, 2017).

Moreover, the collection of large amount medical data with groundtruth is the biggest challenge these

days. The successful deep neural network technique requires thousands of annotated training samples. The annotated medical images with highest level that can be used to train such models of neural networks improving their accuracy.

The purpose of this master thesis is to develop a web application that contains the following main objectives and requirements:

- Build a manual online annotation tool for medical imaging purposes.
- Build an automatic segmentation based on deep learning.
- A user is able to manually correct the semi/automatic annotation tool.
- Able to save groundtruth data for medical image analysis purposes.
- Handle multi users with different roles.
- The uploaded image by a user is stored in a PACS system.

So, the aim of the thesis is to develop a web application called DeepDraw that allows doctors to remotely do above mentioned tasks. We apply the DeepDraw on breast image analysis, more accurately, on segmentation of mass lesions in x-ray images of the breast.

Breast cancer is the most commonly diagnosed cancer among women worldwide after cardiovascular diseases (European Parliament and Council of the European Union, 2017). In 2018, there were over 2 million new cases (Bray F, 2018). For instance, in the European Union it is responsible for one in every six deaths from cancer in women (Luxembourg: Office for Official Publications of the European Communities, 2009). In order to control this alarming mortality rate associated with breast cancer, population screening is recommended by the medical community world wide. Mammography is a widely-used X-ray imaging modality for breast cancer screening as it has the capability to detect various types of lesions such as masses and micro-calcifications (Balleyguier et al., 2005). Among all types of breast anomalies, breast masses are the most frequent, but also the most difficult to detect and segment because of the differences in their size and shape and low signal-to-noise ratio (Dhungel N., 2015). Breast mass morphology is one of the most important characteristic for cancer. The more irregular the shape of the mass, the more likely it is that the lesion is malignant (Oliver et al., 2010).

Although the medical image annotation is quite complex, the medical image analysis needs validation for segmentation. Unlikely, there is no groundtruth or gold standard for the analysis of medical data. In our work, we first started with a typical approach of implementing manual annotation. For this reason, we adopted the web based OHIF[1] medical image viewer platform (Trinity Urban, 2017) to our work. During manual image annotation, most of the time is spent on locating regions of interest by medical expert. In order to speed up the process of image annotation, artificial intelligence is necessary. For that reason, we developed a intelligent annotation tool to precisely segment the breast masses using U-Net convolutional network architecture. Furthermore, a mini PACS system was connected to the platform to upload and store dicom images.

This thesis is organized as follows. Section 2 provides the related work of web based and desktop medical applications. Used datasets, materials, technologies and libraries, the proposed methods for manual and intelligent annotations are described in Section 3. In section 4, experiments are performed on the proposed method and the obtained results are compared with other dataset. In addition, the restriction of the proposed algorithm are explained in Section 5. Finally, Section 6 concludes work of the thesis and suggests some future work.

2. State of the art

Although web applications allow medical experts to share medical images and interpret remote access, still some technological problems that need to be solved. Java is the most popular web technology for developing these applications due to its cross-platform compatibility and remote access (P. T. Looney and Halling-Brown, 2016). Unluckily, Java implementation depends on Java virtual machine (JVM) pre-installation. During installation, certain restrictions, such as administrator permissions, prohibit changes on the computer. In that reason, Java-based web applications do not start due to a failed JVM installation. This flaw leads to many others in web applications. Furthermore, there is a similar issue with ActiveX applications. If this issue remains unresolved, medical experts may not be able to use applications because of the lack of the necessary browser plug-in, which prevents their implementation of remote access interpretations in the future (Qiusha Min and Liu, 2018).

Another popular programming language for the web development is Javascript. When a website uses JavaScript, it is first downloaded to a local user machine and runs locally in the client side by a web browser and relies only on the local hardware, not the network speed. It increases computational power and speeds up the processing of a web site by a web browser. JavaScript works well across different browsers and devices together with advanced features such as HTML5 Canvas, WebGL and CSS3. HTML5 Canvas element allows to

[1] https://github.com/OHIF/Viewers

draw 2D or 3D graphics and can be used for data visualization, animation, game graphics, photo manipulation, and real-time video processing (Basalla, 2014).

It is very critical to choose an appropriate web based medical image viewer to build an application for annotations. Therefore, we explored various web based medical image viewer platforms and reviewed the desktop application for annotations to investigate the state of the art.

2.1. Web Based Applications

Web based applications for annotations in the medical field are limited and they all have been created in recent years. One of the most popular web based app in the medical field is OHIF viewer (Trinity Urban, 2017) provided by the Open Health Imaging Foundation that supports for viewing, annotating, and reporting on DICOM images in 2D images and 3D volumes. It is an open source, built using HTML, CSS and JavaScrip. It can be configured to connect to Image Archives that supports DicomWeb.

SAKE viewer (students, 2017) medical image annotation tool was presented in Capstone project that built based on OHIF viewer. It uses machine learning technologies that allows researchers to label medical images and predict annotation in an automated fashion. However, it is quite sensitive and dependent to the intensity and does not work in complex medical data.

DWV (DICOM Web Viewer) (ivmartel, 2019) was presented by ivmartel (github) that uses Javascript and HTML5 technologies. It is an open source zero footprint medical image viewer library and provides standard tools such as zoom, contrast, drag, draw regions, thresholding and sharpening filters.

Papaya was designed and developed by Jack L. Lancaster, Ph.D. and Michael J. Martinez., it is a pure JavaScript medical image orthogonal viewer that supports DICOM and NIFTI formats, overlays, atlases, GIFTI, VTK surface data and DTI data (Jack L. Lancaster and Martinez, 2019). However, from a design point of view, it makes difficult to construct an automated medical image annotation.

2.2. Desktop Applications

Most of the work on medical annotation has been done in desktop applications. Because those apps run in the server-side and allow any programming languages and libraries. For instance, MITK, ITK-Snap, 3D Slicer enable annotate organs and abnormalities manually or with region growing techniques.

Recently, NVIDIA announced their latest research and technology advances called Clara Train SDK for the Artificial Intelligence (AI) Assisted Annotation and Transfer Learning at the Radiological Society of North America (RSNA) 2018. Clara Train SDK contains of two tools which are Annotation SDK and Transfer Learning Toolkit. NVIDIA AI-assisted Annotation tool enables to accelerate the annotation process with transfer learning in an organ or abnormality. Figure 1 shows how fast it speeds up the process. This can be done simply by clicking a few external points on a particular organ of interest in a 3D medical data and get auto-annotated results for all the 2D slices of that particular organ as illustrated in Figure 3-a. When a medical expert sends the extreme points to the Annotation SDK, a deep learning model receives it as input and returns the predicted results of the segmented organ or lesion (Holger Roth and Roopa, 2019).

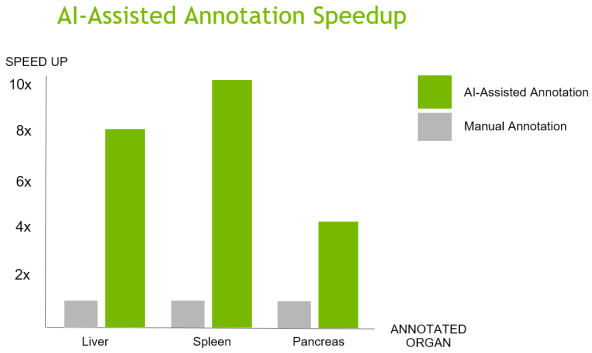


Figure 1: Acceleration of annotations in three organs (*figure from Holger Roth and Roopa (2019)*).

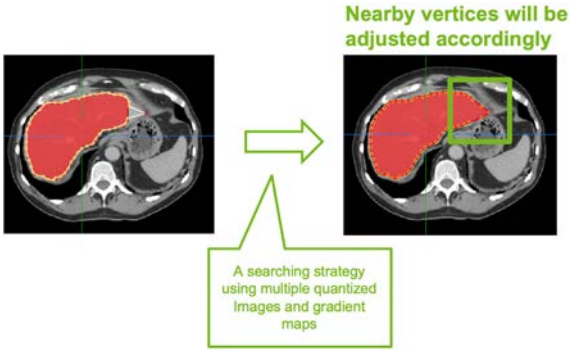


Figure 2: Corrections with smart polygon editing (*figure from Holger Roth and Roopa (2019)*).

This fast process enable scientists to annotate and integrate deep learning tools built into the Annotation SDK with existing medical imaging viewers (e.g. MITK, ITK-Snap, 3D Slicer). This uses a simple application program interface (API) and does not require prior deep learning knowledge. Consequently, knowledge can be increased by analyzing more patient data while still using existing process. Until now, NVIDIA have provided 13 deep learning models for different organs which have been pre-trained on public datasets by NVIDIA. Since deep learning models may give lower annotation accuracy, in this case, 2D smart polygon editing function supports in correcting inaccurate slices. When a medical expert moves a single polygon point on

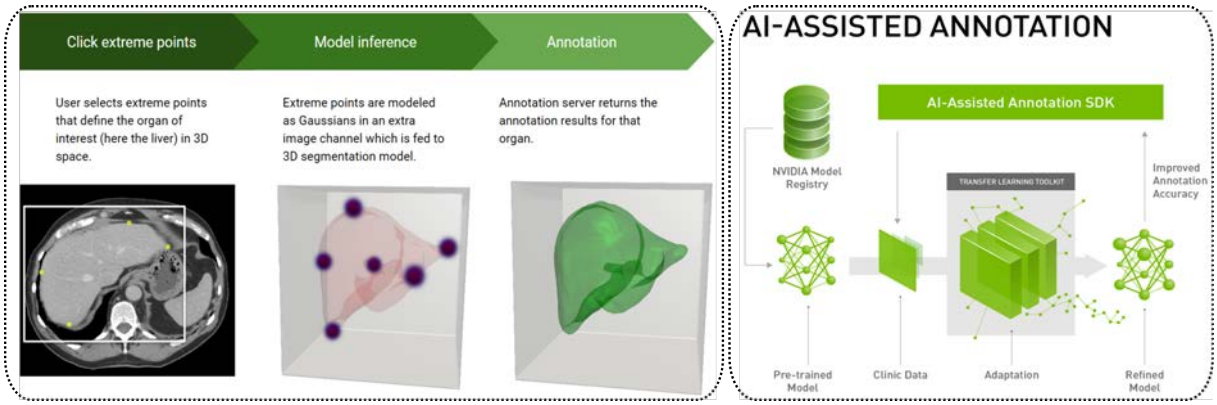


Figure 3: a) Auto-Annotation using extreme points; b) Process of AI-assisted annotation with Transfer Learning Toolkit to improve annotation accuracy (*figures from Holger Roth and Roopa (2019)*).

the 2D slice, all other points around a radius automatically shift to the organ boundaries which is shown in Figure 2.

Obtaining a large well-annotated dataset is much harder in the medical domain, because of the medical expertise that is required to annotate the medical images. In this case, transfer learning is quite suitable way. It makes transfer learning a native for medical image analysis to use pre-trained CNN on larger database and then apply transfer learning to a target field of medical images with limited presence. Figure 3-b demonstrates this process for improving annotation accuracy with Transfer Learning Toolkit.

3. Material and methods

*3.1. **Datasets***

3.1.1. INbreast

The INbreast dataset is full-field digital mammography that consists of 410 images, acquired using MammoNovation Siemens FFDM between 2008 and 2010 at a Breast Centre in a S.Joao Hospital, Porto. This dataset contains examples of normal mammograms, mammograms with masses, mammograms with calcifications, architectural distortions, asymmetries, pectoral muscles and multiple findings. The images were recorded from screening, diagnostic, and follow-up cases in two projections for each breast: craniocaudal (CC) view, which means from head to feet, and a mediolateral oblique (MLO) view, which is a side view, over the lateral 90^{o} projection (Moreira et al., 2011).

The dataset has 115 masses among 107 images, where 90 cases belongs to patients with both breasts, and the remaining 25 cases are from mastectomy patients. The mass lesions were annotated accurately by a medical expert According to BI-RADS, a mass is defined as a three-dimensional structure showing convex outer boundaries, usually appearing on two orthogonal views. The average mass area is 479 mm^2 with a standard deviation of 619 mm^2 and the area of masses are from 15 mm^2 to 3689 mm^2. In addition, the dataset is 16-bit images in Dicom format, with two matrix size 2560 x 3328 and 3328 x 4084 pixels.

In this work, above mentioned 107 images of masses were used to experiment and evaluate the investigated method on segmentation problem.

3.1.2. OMI-DB

The Optimam Mammography Image Database (OMI-DB) is a comprehensive that contains of unprocessed and processed digital mammography images. The database were created support medical imaging research. The images were acquired from different mammogram equipment systems such as GE, Siemens and Hologic. The groundtruth of this database was made a rectangular shape around the lesions. The dataset was used to test our trained model in INbreast dataset, that to know how it behaves in other datasets, despite the different intensity.

*3.2. **Medical Image Viewer***

Building a web application from scratch is quite complicated and it takes a lot of time and effort in terms of design and architecture of the application. As mentioned in an introduction section, the medical image viewer platform was adopted into our work. In this section, we describe the library and platform used for displaying medical images.

3.2.1. Cornerstone JS

Cornerstone[2] is a lightweight JavaScript library for displaying medical images that provides a complete web based medical imaging platform. It is not a full application, but can be used as a component to build medical imaging applications. Nowadays, Cornerstone is very popular in medical imaging and many companies

[2] `https://github.com/cornerstonejs`

use this library in their application production. It enables HTML5 canvas element latest technology in modern web browsers including tablet, mobile and desktop. The Cornerstone core is an independent of the actual container used to store the image pixels and transport mechanism used to obtain an image data. Actually, Cornerstone Core itself does not have the ability to read and parse or load images and alternatively, it relies on one or more image loaders to work with, i.e. CornerstoneWADOImageLoader. The purpose of doing this is to avoid constraining to work within a single container and transport. It allows to load images with highest performance image display from any type of image container using any variety of transport and it doesn't require the conversion to an alternative container or transport. Cornerstone Image Loader refer over HTTP (WADO-URI) or DICOMWeb (WADO-RS) that can communicate with PACS systems or Dicom servers. So, the Cornerstone enables to display a 8 or 16 bit grayscale and RGB color medical images as well (Hafey, 2014).

The cornerstone provides the main following features:

- Resize - changes a width or height of the image
- Change slice - shows different an image slice, e.g. in 3D MRI
- Window level - adjusts brightness or contrast
- Zoom and pan - zooms an image and moves view
- Interpolation - turns on/off interpolation
- HTML overlays - illustrates overlays on top of the image using HTML
- Event handling - if the top of the image is changed, image is updated everytime
- Multimage - shows two medical images on one page
- Flip and rotate - flips a image vertically or horizontally, and rotates clockwise or anti-clockwise
- WebGL - renders 2D and 3D graphics
- False color mapping - creates a false color mapping (i.e Hot Iron, HSV, Gray etc.)
- Display area - displays an image any area of the page

Moreover, the Cornerstone provides the following additional tools to process reports, measures and annotations on an image:

- Angle tool - defines an angle
- Arrow annotate - marks the reports with arrow
- Bidirectional - measures a bidirectional length
- Cobb angle - defines a cobb angle
- Elliptical ROI - measures the area of the elliptical ROI
- Freehand Mouse - draws a contour and measures an area of the contour
- Length - measures a length of the two points
- Probe - determines a intensity of the pixel and coordinates of the given pixel
- Rectangle ROI - measures the area of the rectangular ROI
- Magnify - zooms given particular area
- WWWC Region - changes a contrast and brightness in accordance with the specified region

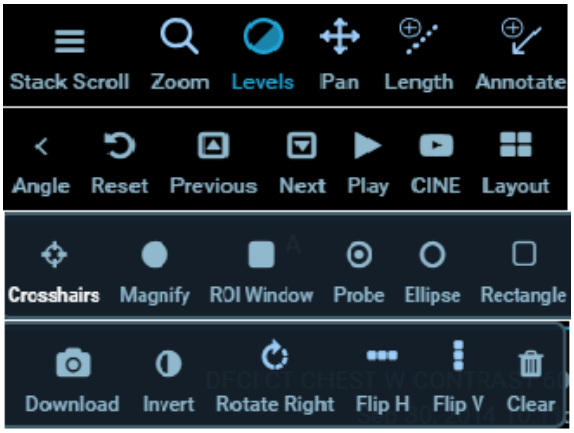


Figure 4: Cornerstone and OHIF viewer tools.

3.2.2. OHIF Medical Image Viewer

The Open Health Imaging Foundation (OHIF)[3] Viewer is an open source, web-based, medical imaging viewer platform. The application platform was created using JavaScript, HTML, CSS, and displays based on Cornerstone library. It is built in Meteor full-stack web framework that consists of three types of applications: the OHIF Viewer, Lesion Tracker and the Standalone Viewer.

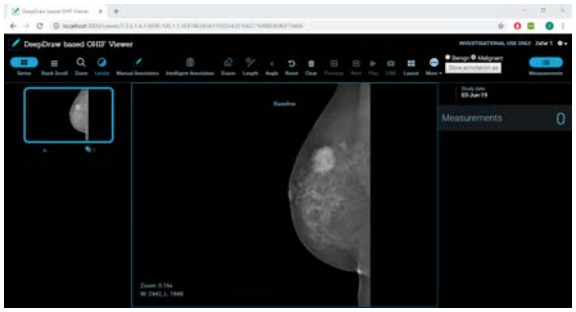


Figure 5: OHIF viewer web application in the browser.

The OHIF Viewer is a universal implementation of the OHIF platform intended for users of general radiology that implements most of the features of the cornerstone. One of the main functionality is that it can be set up to connect to PACS systems that supports DicomWeb Standard. Furthermore, the maintained extensions add support for viewing, reporting, annotating on 2D and 3D DICOM images which are shown in Figure 4.

[3] www.ohif.org

The LesionTracker is second example implementation of the OHIF platform that focuses on oncology progress and partially funded by the National Cancer Institute (NCI). This implementation is targeted to measure and follow the lesions, and to store the measurements of overlay images into a database. It is suitable to compare and analyze the images in two or more windows but the platform is not enriched with various tools.

The StandaloneViewer enables only the client-side portions of the OHIF Viewer (single-page viewer), without the image managing feature. It can be used for deploying as a stand-alone web application.

As a starting point for the development of more reliable, full-featured viewer applications, we decided to build our annotation tools on top of the OHIF Viewer due to its performance and general purpose. The Figure 5 illustrates the main web page of the OHIF viewer in the browser. The Figure 6 shows the study list of OHIF viewer that images can be managed in this page including searching by patient name, study date, modality, study description, retrieving.

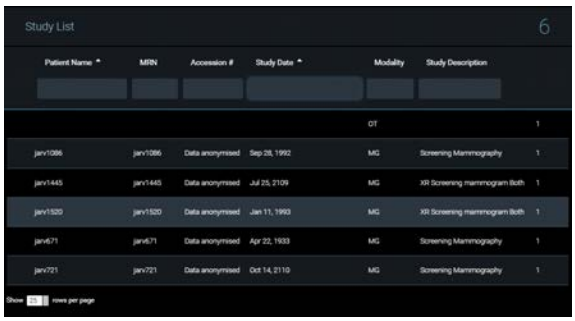


Figure 6: Study list of OHIF viewer.

3.3. *PACS System*

The Orthanc[4] is open-source software that provides powerful Dicom server for clinical and medical research. It is a research work by Sebastien Jodogne from the University Hospital of Liege (Belgium), which is currently being developed and maintained by Osimis S.A. The Orthanc works in any computer running Windows, Linux or Mac OS X to store dicom images that performs task as a mini PACS system. A picture archiving and communication system (PACS) is a medical imaging technology which enables secure storage and image transmission to multiple machines (Svb et al., 2018).

One of the main characteristic of the Orthanc is that it provides a RESTful API and supports Dicom standart. The Orthanc has a plugin mechanism for adding new modules such as a DicomWeb, web viewer, PostgreSQL and MySQL database back-ends that expands the capabilities of its REST API (Jodogne, 2018). This API also allows full CRUD operations (create, read, update and delete) on the dicom data and enables the following DIMSE Service (TCP/IP):

- C-Echo - test the connection between two devices
- C-Store - send images from the local imaging device to a remote device
- C-Find - search the content of a remote device
- C-Move - retrieve images from a remote device

The Figure 7 shows the web user interface of the Orthanc and named Orthanc Explorer that listens on the port 8042.

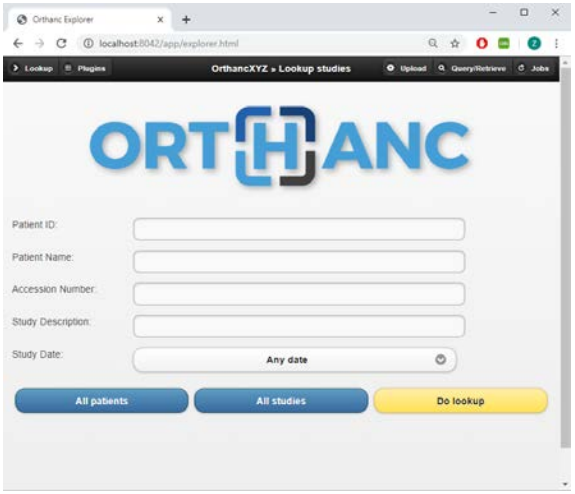


Figure 7: Orthanc web user interface.

The DicomWeb is the web protocol (i.e HTTP or HTTPS) that provides a simple mechanism for accessing a dicom data and contains the following a set of RESTful services:

- WADO-URI - retrieve single dicom instances
- WADO-RS - retrieve dicom objects
- QIDO-RS - search for dicom objects
- STOW-RS - store dicom objects

One good illustration for this example is shown in Figure 8, where HTTP get request is sent to WADO-RS RESTful API by a user-side and Dicom Service response a specific image study which was requested by a user and QIDO-RS also searches and response information to user respectively.

In our case, a user is the OHIF medical image viewer application and dicom service is the Orthanc. As mentioned in section 3.2.2, the OHIF viewer also enables DicomWeb Standard for exchanging medical images and its metadata. Since, the OHIF viewer application works locally in our machine, we connected it to Orthanc Dicom server by configuring the following ports and urls:

- WADO-URI: http://localhost:8042/wado
- QIDO-RS: http://localhost:8042/dicom-web
- WADO-RS: http://localhost:8042/dicom-web

[4] `www.orthanc-server.com`

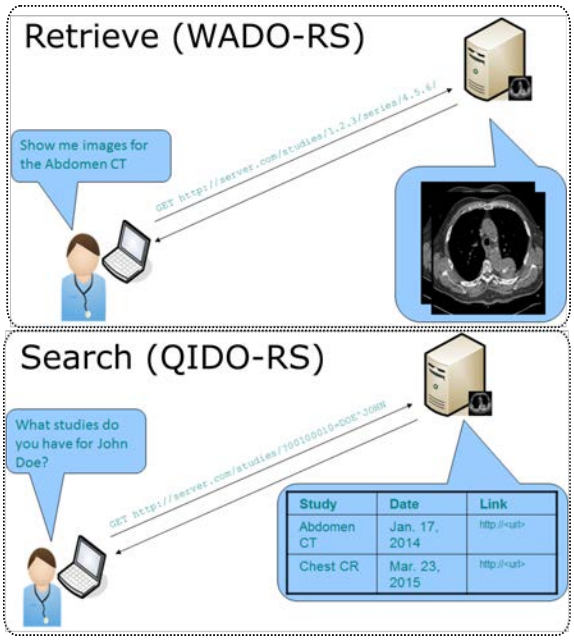


Figure 8: Search (QIDO-RS) and Retrieve (WADO-RS) DicomWeb Services (*figures from dicomstandard.org*).

3.4. *Authentication*

When a web application is used in the medical field, all patient data must be safe and protected from others. In order to avoid this issue, it is necessary to include a authentication functionality into the web app. It also makes easy for each doctors to remotely access and register. In fact, that the authentication functionality was developed only for the Lesion Tracker viewer but not for the OHIF viewer. We integrated this functionality and installed the package into the OHIF viewer as well. Originally, the used authentication package has been implemented by Clinical Meteor[5] named clinical-entry to use in Meteor frameworks. So, the package provides 4 pages for clinical web apps: "Sign In", "Sign Up", "Forgot Password", "Change Password", and "Logout". Its components and routes were embedded with MongoDB database similarly as Lesion Tracker viewer. It allows a token based authentication to secure a password and a user information is saved in MongoDB database. We did not connect the "Forgot Password" implementation into web platform. Because it requires an automatic response. It can be implemented during deployment process of web app into hosting provider. The Figures 9 and 10 show the registration and sign in process.

3.5. *Manual Annotation Implementation*

This section describes the adding of two functionalities to the OHIF viewer which are manual annotation and eraser tools.

[5] `https://github.com/clinical-meteor/entry`

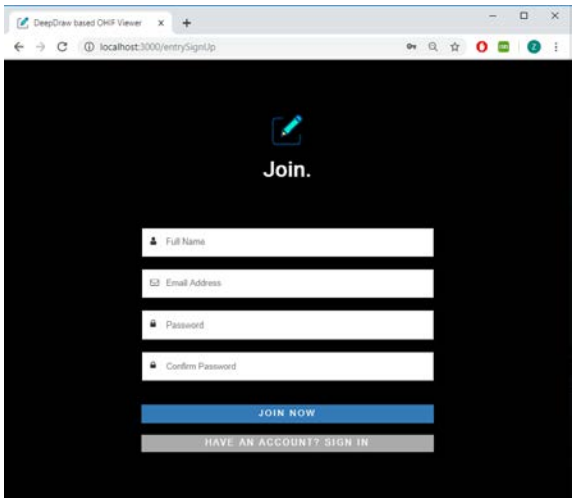


Figure 9: Registeration page.

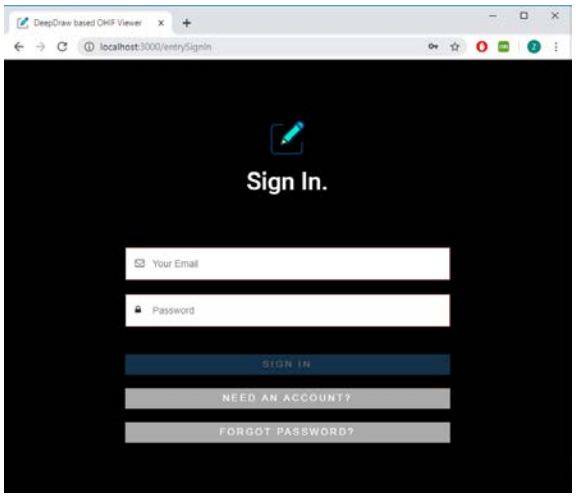


Figure 10: Sign in page.

3.5.1. *Manual Annotation Tool*

We started our work with a typical approach, that is a manual annotation. For this reason, "Freehand Mouse" tool from the Cornerstone javascript library was used and incorporated into the OHIF viwer platform. This tool allows to draw on the image with two options. First option is to draw point by point where a user is able to add points sequentially on the image, a tool adds a line between two point and displays on the screen after each point. Second option similar to normal pencil that a user can draw by moving a mouse continuously. While moving mouse, a tool adds points and represents similarly as a first option (i.e connected points with lines). This can be done by holding "shift + mouse left button" in the keyboard. However, in both cases, a user must stop a drawing by connecting first and last points where a user started and ended. The Figure 11 illustrates both drawing options.

After applying this tool, region of interest (ROI) is obtained. The tool also is able to show the mean and

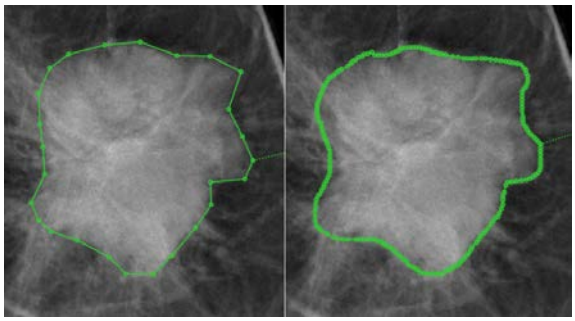


Figure 11: Example of manual annotation in two option (point by point and drawing with many points).

standard deviation of pixels intensity of the ROI and area of the ROI (pixels^2). This functionality can be helpful for doctors or researchers. One of the main feature of this tool is also a removing points of the contour (at least 3 points must be left to display a contour), adding points between two lines and changing coordinates of any points. The adding and removing can be done by holding "ctrl + mouse left button" in the keyboard.

The pencil icon was designed for the manual annotation tool button (see Figure 12) and, configured and registered into the toolbar of the OHIF viwer.

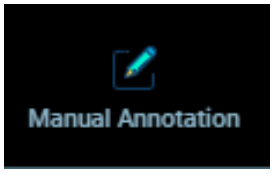


Figure 12: Manual Annotation tool button.

3.5.2. Eraser Tool

The OHIF viewer provides "Clear" tool which cleans everything on the image (i.e. drawings, annotations, measurements, lines, marked reports). But there is no functionality to remove the annotations one by one. As an additional tool, we integrated "Eraser tool" from Cornerstone javascript library to remove the annotations if they are not correct. The implementation was done in the same way as manual annotation tool (see Figure 13). This tool works by clicking right mouse button on unnecessary item.

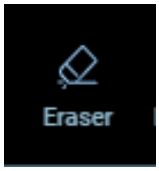


Figure 13: Eraser tool button.

*3.6. **Storing Annotation***

After annotation of lesion or organ by a medical expert, it is necessary to collect it in a safe place. So, in this section, we introduce the implementation of the storing annotation. We decided to save the annotations in a database as well as in a local storage. The annotated ROI contains of points that represent coordinates. It is not easy task to generate the groundtruth of annotated image in javascript. Therefore, the implementation was divided into two tasks on the client-side and on the server-side.

3.6.1. Design (client-side)

In the client-side (i.e the OHIF viewer), we designed the "Store annotation" button in the form of radio button that a user can select the type of annotated lesion. Since we are working on mammography images, we included benign and malignant options which is shown in Figure 14.

Figure 14: Annotation storing button.

3.6.2. Storing Annotation REST API (server-side)

In order to generate and save a groundtruth of annotated image, we built a new server on the Flask framework. The Flask is a lightweight and powerful micro web framework which uses Python language. In this framework, the REST API was created to do all task including generating and storing process. The REpresentational State Transfer (REST API) is a software architectural style that can handle requests and receive responses via HTTP protocol such as GET, POST, PUT, DELETE. We used secure POST method to transfer data from client (OHIF viewer) to server (Flask) in HTTP.

After annotating and selecting a type of lesion, when a user clicks the store annotation radio button in the web browser, the annotation is saved in the following structural order (see Figure 15):

1. The client-side queries to the Orthanc dicom server (port: 8042) to get a image information using WADO protocol.

2. The client-side retrieves a response from the Orthanc dicom server

3. The client-side gets the name, size and view of the image. Then it generates the coordinates (x,y) of the points of annotated ROI. The name, size and view of the image, coordinates and lesion type are converted into one JSON format data and it will be ready for the next step. JSON (JavaScript Object Notation) is a data-interchange format that is easy for machines to parse and generate as well as for people to read and write.

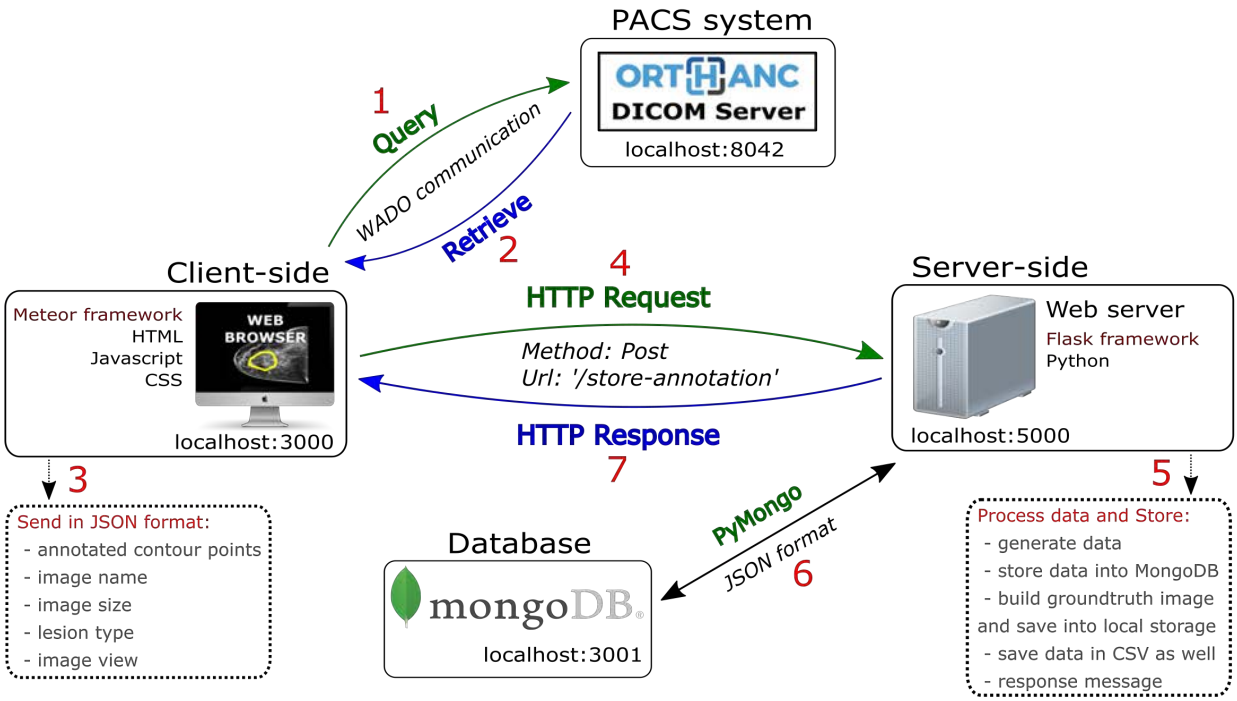


Figure 15: Structure of the storing annotation REST API.

4. The client-side sends JSON data to the server-side (port: 5000) with HTTP POST request. This was done using Ajax technology that enables to send and receive data from the server asynchronously, without affecting the display and behavior of the existing page. In the server-side, "store-annotation" REST API was created with POST method that it receives the JSON data (in url: http://localhost:5000/store-annotation) and performs the next processes.

5. The REST API is in Python language that first generates all data from JSON format. Then it builds a groundtruth of the annotated image using coordinates of the points of annotated ROI and image size. This was done with wonderful fillPoly OpenCV function. The API is able to build any number of annotated contours. A groundtruth of the image is saved into local storage with original name and size in png format as well as all the annotated information in CSV format, so later that it can be used by researchers.

6. All the annotated information is stored in the MongoDB database in JSON format so later it can be used again for visualizing in the OHIF viewer.

7. Finally, in the last step, the REST API responses to the client-side with message that storing process was performed successfully.

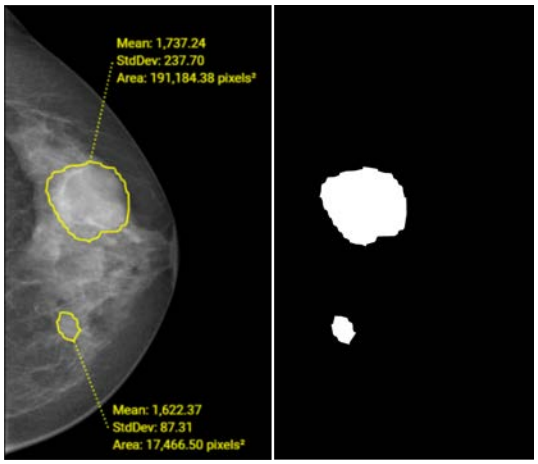


Figure 16: Manually annotated image and its stored groundtruth.

3.7. *Intelligent Annotation Implementation*

During the manual image annotation, most of the time is spent on locating regions of interest by medical expert. In order to speed up the process of image annotation and segment the breast masses on mammography images, we implemented an intelligent annotation tool as a second prediction of annotation from a machine. So, in this section, we introduce our implementation of intelligent tool.

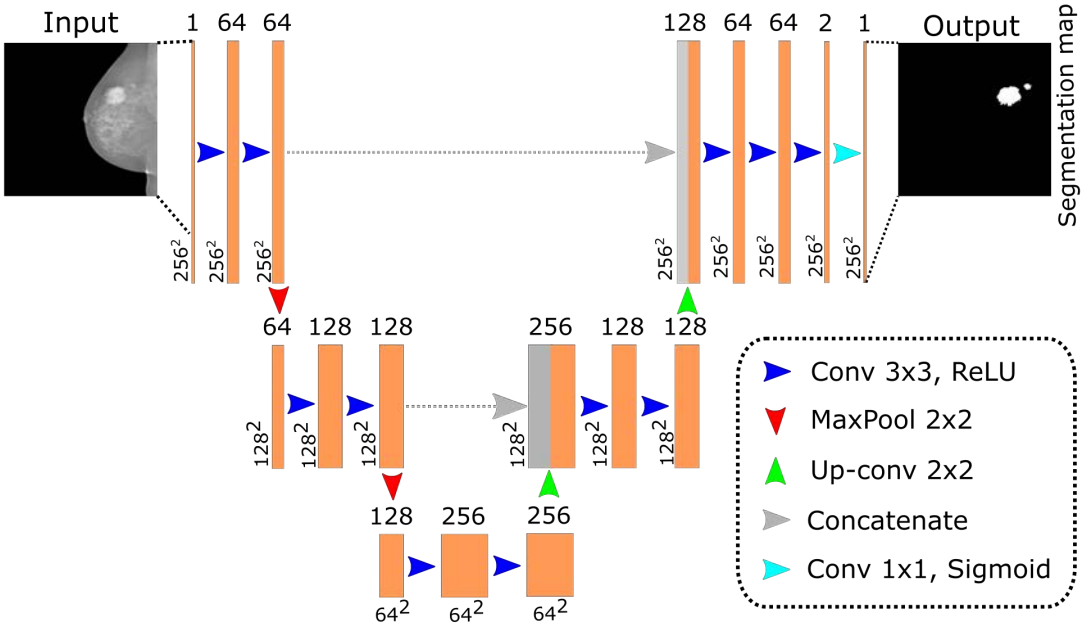


Figure 17: U-Net architecture.

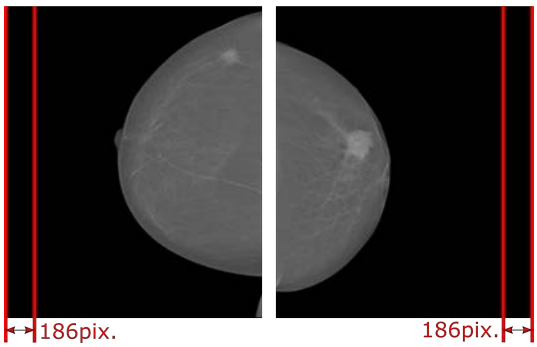


Figure 18: Cutting 186 pixels of black side from left and right.

3.7.1. *Data Preparation*

Before training our network, we applied preprocessing techniques on the 107 mammogram images containing mass lesions from the INbreast dataset. The mammogram images consists of two different matrix size (2560 x 3328 and 3328 x 4084 pixels). To make them same size, we cut 186 pixels on the black side of the 3328 x 4084 pixels image from left and right according to the breast side (see Figure 18). Then, the cropped images were resized to 2560 x 3328 pixels. The reason of cropping 186 pixels is that it preserves its aspect ratio according to the image resizing and also black side is not needed for network to train. The images were converted from 16 bit to 8 bit and rescaled between 0 and 255.

3.7.2. *U-Net Network Architecture*

In this work, our training dataset is too small. Therefore, the well known U-Net architecture was used. The U-Net is a Convolutional Neural Network (CNN) based segmentation algorithm originally proposed by (Ronneberger et al., 2015) for biomedical image segmentation. The U-Net includes a contractive downsampling and expansive upsampling path with skip connections between the two parts, which uses standard convolutional layers. The main advantage of using this architecture is that it performs well with limited training data by concatenating multi-resolution information. After trying a few experiments, a small design changes were made to the standard U-Net implementation. It should be noted that a groundtruth of the image disappears after each layer. Therefore, layers of architecture were reduced. The typical U-Net was designed for images of size 572 x 572. The input layer dimension of our architecture was modified to 256 x 256. All the convolution layers uses ReLU activation functions with kernel size of 3 x 3 and ”he” normal kernel initializer, except the last layer which uses sigmoid activation function with kernel size of 1 x 1 to define an output probability map of two classes as shown in Figure 17. There are two max pooling layers which reduces the size of feature maps by 2. After each up-sampling and convolution with 2 x 2 kernel size, it is concatenated with previous feature map of layers. The network contains total of 1,862,789 parameters.

3.7.3. Network Training

The dataset consists of a total of 107 images with mass lesions. To test fairly our training network, the dataset was randomly divided into 80 % (85 images) training and 20 % (22 images) testing.

The U-Net was trained with Adam optimizer, as proposed by (Kingma and Ba, 2015). The method is an adaptive learning rate optimization algorithm that computes individual adaptive learning rates for different parameters using estimations of first and second moments of the gradients. In our work, an Adam optimizer was maintained with a learning rate of 0.0001.

A loss function was defined with a binary crossentropy during training the network for each iteration (see Eq. 1).

$$BCE = -\frac{1}{N}\sum_{i=0}^{N} y_i \cdot log(\hat{y}_i) + (1 - y_i) \cdot log(1 - \hat{y}_i) \tag{1}$$

where y is the label (1 for mass lesion and 0 for no mass) and p(y) is the predicted probability of the pixels being mass for all N pixels. It adds log(p(y)) to the loss for each mass pixel (y=1), that is the log probability of it being mass. It adds log(1-p(y)) vice versa that is the log probability of it being no mass for each mass pixel (y=0).

As a evaluation metrics, the Dice similarity coefficient was calculated for training and testing the network between two sets (see Eq.2).

$$DSC = \frac{2 * |X \cap Y|}{|X| + |Y|} \tag{2}$$

where X is a groundtruth and Y is a prediction.

The proposed method was developed on Python using Keras (Chollet et al., 2015) with Tensorflow backend. The implementation was developed on a 64-bit Ubuntu operating system using a Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz and Nvidia Tesla K40c GPU with 12 GB of video RAM. During the network training, CS Vlogger Keras callback functions were used to store the values of the loss functions and metrics for each epoch. The Model checkpoint callback was used to store the weights of the network after every epoch as a best, if there is a decrease in the loss function.

*3.8. **Intelligent Annotation Deployment***

This section describes the process of deep learning deployment for the production purposes using our pretrained model on segmentation breast masses and shows how Intelligent Annotation is installed into the OHIF viewer web application.

3.8.1. Design (client-side)

In the first stage, an icon was designed for the intelligent annotation tool button as a brain in the OHIF viewer (see Figure 19). Then, the button was configured and registered into the toolbar of the OHIF viwer like a similar manual annotation button.

Figure 19: Intelligent Annotation tool button.

3.8.2. Intelligent Annotation REST API (server-side)

The Intelligent Annotation tool was designed similarly as a storing annotation architecture building REST API. Our pre-trained deep learning model was developed on Python using Keras with Tensorflow backend. So, we are able to run this model in Flask framework for segmenting breast mass lesions. In order to do this, new REST API was created named ”segmentation” with POST method and it can be seen in URL format follows: ”http://localhost:5000/segmentation”.

If we follow the order structure (see Figure20), it will be easy to understand the logic of the algorithm. So, when a user clicks the intelligent annotation button in the web browser, the computer-assisted annotation gives its prediction in the following structural process:

1. The client-side queries from the Orthanc dicom server (port: 8042) to get a dicom image using WADO protocol.

2. The client-side retrieves a response from the Orthanc dicom server.

3. The client-side gets pixel data from a dicom image and, its height and width of image from a dicom tag. It is necessary to have an image size to build image from pixel data. Because, in javascript, a retrieved image pixel data looks like an array object in one line, not numpy. Then, the OHIF viewer converts the pixel data and image size to JSON string.

4. The client-side (port: 3000) sends JSON data to the server-side with HTTP POST request (URL: http://localhost:5000/segmentation) using Ajax to able to get prediction result and visualize for a user.

5. In the server-side, the new created REST API receives the JSON data and performs the follows:

    - First, 16 bit image is generated and built using reshaping. Then, 16 bit image is converted to 8 bit and rescaled between [0, 255].

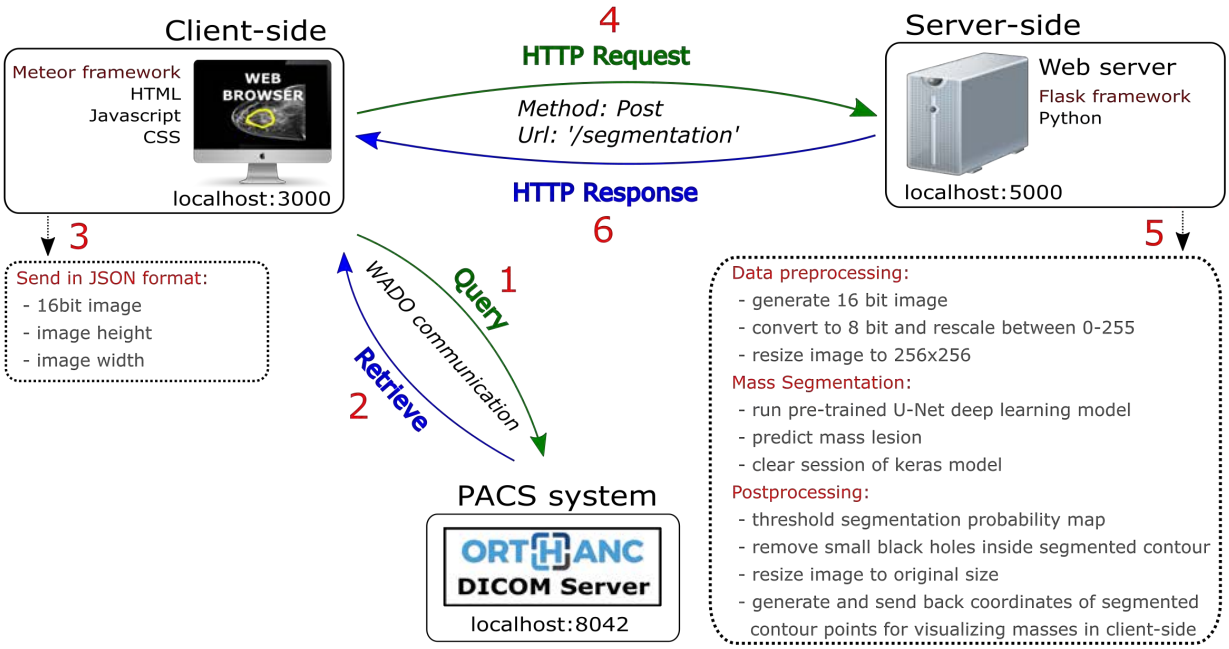


Figure 20: Structure of the Intelligent annotation REST API.

The converted image is resized to the size of input (256x256) of U-Net network.

- The pre-trained U-Net deep learning model is run as usual (i.e as how we test it for prediction, (1) test generator generates image for keras model, (2) network and model weight is loaded, (3) predict generator gives segmentation probability map). After that, a API clears a session of keras model. It is significant that a session is cleared, otherwise in the next predict, it may give an error due to keras tensors.
- Then, applied thresholding (0.5) for a segmentation probability map of image. The small black holes are removed if there are inside a segmented contour using closing morphological transformations. A segmented image is resized to original size and coordinates of segmented contour points is generated using openCV findContours function.

6. This contour points of segmented image are required to visualize in the client-side (web application) for a user. Finally, API returns the contour points to the OHIF viewer for visualizing.

The OHIF viewer can visualize any number of segmented contours. One of the main advantage in this method is that a user is able to correct the segmented contour if image is not segmented well. This will be done using manual annotation tool.

4. Results

This section provides the results of the both annotation methods in different datasets, training and testing experiments.

4.1. Manual annotation and Execution time

By testing the manual annotation tool, it was noted that a medical expert is spent approximately one minute per image for annotating mass lesion or more depending on the complexity. The tool for storing annotation works very quickly, almost one second per image. A communication between the Orthanc PACS system and OHIF viewer is also fast.

4.2. Training results

Figure 21 shows the graphs for the training loss function and dice coefficient of U-Net for 500 epochs. It can be noticed that the dice coefficient was reached significantly to 0.95 in 200 epochs and gradually to 0.99 including decreased overtime. The loss function was suddenly droppped at the beginning of training and then decreased slowly.

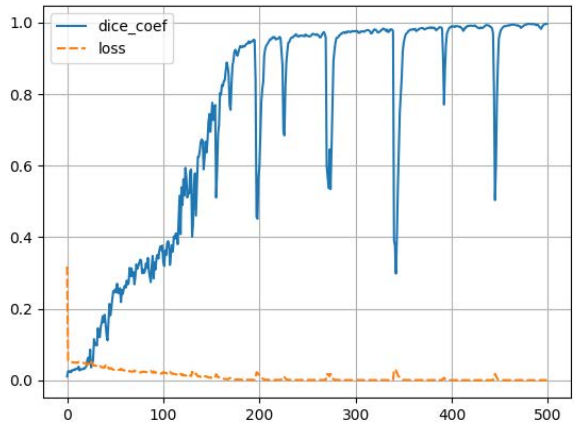


Figure 21: Training dice coefficient and loss.

4.3. Testing results

The box plot Figure 22 shows that the average dice coefficient of 0.52 was obtained for 22 test images in INbreast dataset. Two images in test dataset were obtained with almost close to 0. This and other different factors have played essential role in making it lower. When a dice is calculated between groundtruth and predicted image, if they do not intersect, a dice will be obtain with 0.

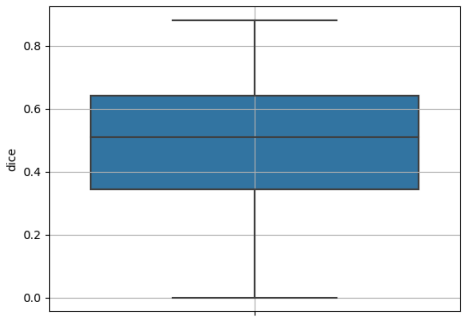


Figure 22: Dice coefficient of test dataset (22 images).

The Figure 23 displays the overlap with a good, medium and poor of test results. Where yellow color is the groundtruth and along it, predicted image is specified with red color.

4.4. Intelligent annotation

The intelligent annotation tool was tested in different datasets. The trained images were applied in the OHIF viewer web application that the Figure 24 shows the example before and after segmentation by intelligent annotation tool. The figure also shows the mean, standard deviation and area of the segmented mass. This intelligent annotation tool took approximately 30 seconds for all process. It should be noted that it was tested on the CPU Intel Celeron N2930@ 1.83GHz without GPU.

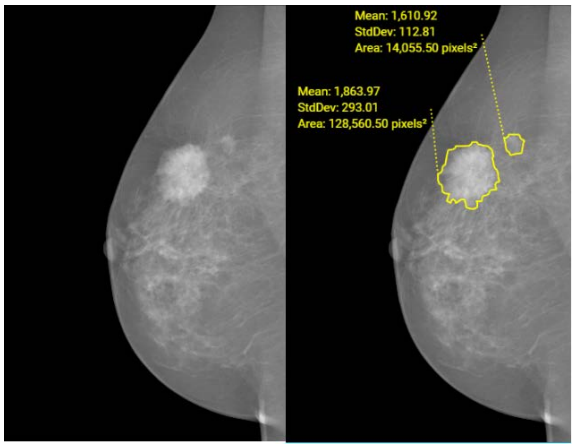


Figure 24: Example of the result of the intelligent annotation tool.

5. Discussion

In this section, the results of the algorithms implemented are discussed and the conclusion of this work are given.

5.1. Manual Annotation

The medical data is quite complex, even it is difficult to distinguish a medical tissue. A machine does not always help to solve any problems on image annotation. For that reason, our manual annotation tool can be useful for the correction of predictions of machines. It does this job fast and allows to work remotely for a medical expert.

5.2. Intelligent Annotation

As observed on the previous section, the results of dice coefficient of the test data is too low than the training dice, it seems that the network is over-fitted with the training data. However, in terms of the number of trained images and large full mammograms, the results are a normal for this task.

In this thesis, we were more focused on the deployment part of the web application. Furthermore, the proposed method was tested on the Optimam dataset using intelligent tool in the web app. This dataset is completely different in the term of intensity and does not provided an accurate groundtruth which has only a groundtruth of rectangular masses boundary. The results shows that the trained-model is able to segment masses in some way (see Figure 25). The yellow contours are prediction by U-Net model and the red color contours are groundtruth that was annotated by us.

As mentioned previous section, all process of intelligent annotation was performed in 30 second. Despite

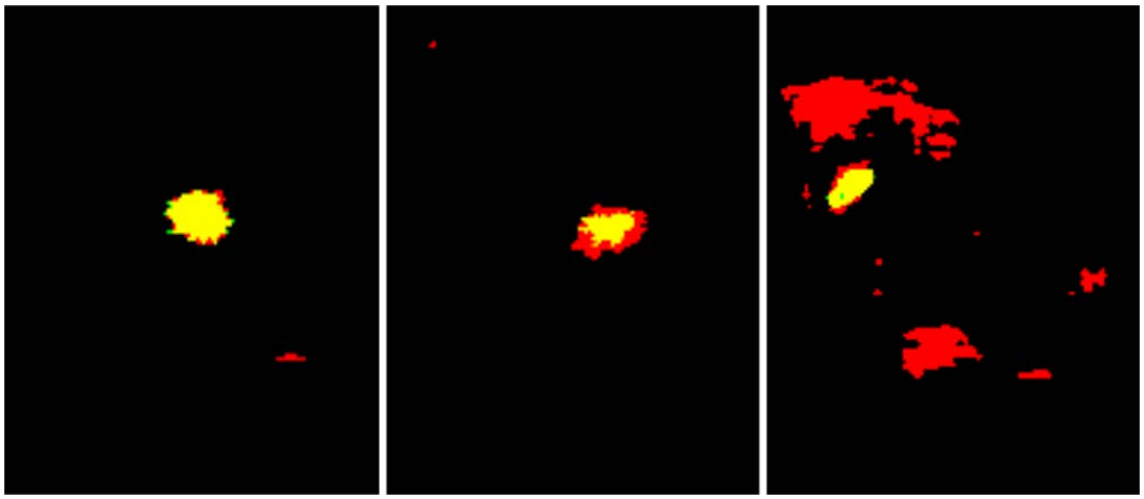


Figure 23: Example of testing segmentation results in INbreast (good, medium, poor), yellow = groundtruth, red = predicted.

the fact that the computational time of the intelligent annotation can be achieved great performance with a GPU machine and powerful computer.

As a observed results, we realized that the better approach can be performed starting from the implementation of detection of masses then can be applied segmentation algorithm. It is important to point out that the deployment of the deep learning based method should be done in auto mode where an image should be tested immediately when a user uploads an image to the PACS system. This can save time of medical experts.

6. Conclusions and Future Work

In this thesis work, we developed the DeepDraw deep learning based tool that is able to segment breast masses and correct it in the web medical image application. In general, our proposed method has performed well in terms of development of manual and intelligent annotation and web application with a wonderful interface possibility. This new approach can contribute and serve for the computer-assisted web application in the medical field as well as reduce time and effort on annotation for medical expert. Since, we have not yet deployed the web application into a host due to time limit.

Some future works can be added into the web application. For example, creating a web page to able to manage with a stored annotation information such as read, delete, visualize. The DeepDraw can also be implemented for other medical segmentation problems including 3D volume image. Using strategy of immediately predicting uploaded image can be helpful to visualize a result and save time of medical expert. Current Docker and Kubernetes software technologies makes a web applications powerful on deployment and in all respects. The Google Cloud Healthcare API can be a powerful option for storing medical imaging data in the cloud that the Google Cloud software is the service provider as a PACS that enables cost effective image storage solution.

An alternative to deploying your own PACS is to use a software-as-a-service provider such as Google Cloud. The Cloud Healthcare API promises to be a scalable, secure, cost effective image storage solution for those willing to store their data in the cloud. It offers an almost-entirely complete DICOMWeb API which requires tokens generated via the OAuth 2.0 Sign In flow. Images can even be transcoded on the fly if this is desired. The Cloud Healthcare API is a very attractive option because it allows us to avoid deploying the Meteor server entirely. We can just deploy OHIF as a client-only static site application.

7. Acknowledgments

Firstly, I would like to thank all my supervisors Robert Marti and Oliver Diaz for their support and guidance and for giving opportunity to work on the development of web application. I would also like to thank MAIA for providing Erasmus Mundus scholarship and other members of the VICOROB Lab for their sharing knowledge and guidance.

References

Akkus, Z., Galimzianova, A., Hoogi, A., Rubin, D., Erickson, B., 2017. Deep learning for brain mri segmentation: State of the art and future directions. Journal of Digit Imaging 30, 449–459. doi:10.1007/s10278-017-9983-4.

Balleyguier, C., Kinkel, K., Fermanian, J., Malan, S., Djen, G., Taourel, P., Helenon, O., 2005. Computer-aided detection (cad) in mammography: Does it help the junior or the senior radiologist? European journal of radiology 54, 90–6. doi:10.1016/j.ejrad.2004.11.021.

Basalla, D., 2014. Browser-based Medical Image Viewer using WebGL. Master's thesis.

Bray F, Ferlay J, S.I.S.R.e.a., 2018. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36

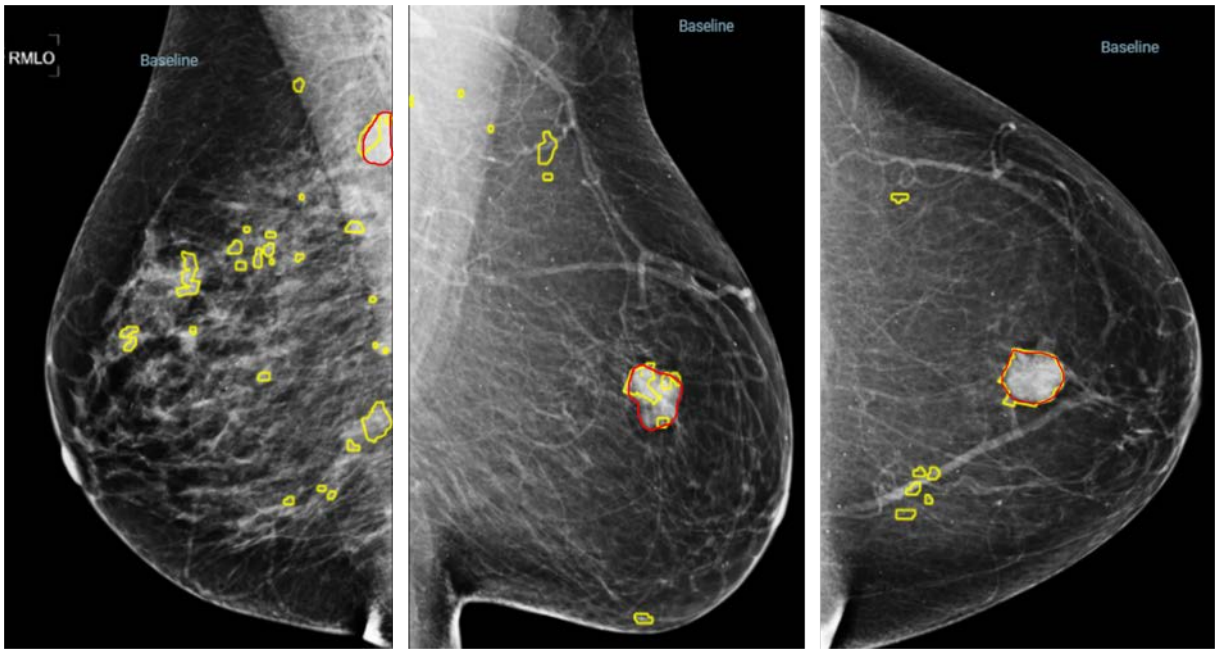


Figure 25: Example of predictions of the Intelligent Annotation tool in the Optimam dataset.

cancers in 185 countries. CA: a cancer journal for clinicians , 394–424doi:10.3322/caac.21492.

Chollet, F., et al., 2015. Keras. https://keras.io.

Dhungel N., Carneiro G., B.A., 2015. Deep learning and structured prediction for the segmentation of mass in mammograms. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 9349, 605—-612. doi:https://doi.org/10.1007/978-3-319-24553-9_74.

Eli Gibson, Wenqi Li, C.S.L.F.D.I.S.e.a., 2017. Niftynet: a deep-learning platform for medical imaging. Comput Methods Programs Biomed , 113–122doi:10.1016/j.cmpb.2018.01.025.

European Parliament and Council of the European Union, 2017. Breast cancer facts and figures 2017-2018 .

Hafey, C., 2014. Cornerstone js.

Holger Roth, P.C., Roopa, M., 2019. Fast ai assisted annotation and transfer learning powered by the clara train sdk. URL: https://devblogs.nvidia.com/annotation-transfer-learning-clara-train/.

ivmartel, 2019. Dicom web viewer. URL: https://ivmartel.github.io/dwv/.

Jack L. Lancaster, P., Martinez, M.J., 2019. Papaya. URL: http://rii.uthscsa.edu/mango/index.html.

Jodogne, S., 2018. The Orthanc ecosystem for medical imaging. Journal of Digital Imaging 31, 341–352. URL: https://doi.org/10.1007/s10278-018-0082-y, doi:10.1007/s10278-018-0082-y.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization, in: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. URL: http://arxiv.org/abs/1412.6980.

Luxembourg: Office for Official Publications of the European Communities, 2009. Health statistics atlas on mortality in the european union. International series of monographs on physics .

Moreira, I., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M., Cardoso, J., 2011. Inbreast: Toward a full-field digital mammographic database. Academic radiology 19, 236–48. doi:10.1016/j.acra.2011.09.014.

Oliver, A., Freixenet, J., Martí, J., Perez, E., Pont, J., Denton, E., Zwiggelaar, R., 2010. A review of automatic mass detection and segmentation in mammographic images. Medical image analysis 14, 87–110. doi:10.1016/j.media.2009.12.005.

P. T. Looney, K.C.Y., Halling-Brown, M.D., 2016. Medxviewer: providing a web-enabled workstation environment for collaborative and remote medical imaging viewing, perception studies and reader training. Radiation Protection Dosimetry 169, 32—-37. doi:10.1093/rpd/ncv482.

Qiusha Min, Z.W., Liu, N., 2018. An evaluation of html5 and webgl for medical imaging applications. Journal of Healthcare Engineering 2018, 11. doi:10.1155/2018/1592821.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. CoRR abs/1505.04597. URL: http://dblp.uni-trier.de/db/journals/corr/corr1505.html#RonnebergerFB15.

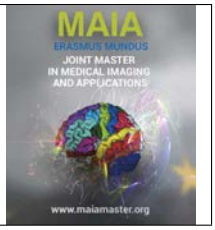
students, C., 2017. Sakeviewer. URL: http://sakeviewer.com.

Svb, D., Deshmukh, V., Mohan Kulkarni, D., Shailesh Kolharkar, M., 2018. Pacs: An overview of the technology and related issues.

Trinity Urban, Erik Ziegler, R.L.C.H.C.S.e.a., 2017. Lesiontracker: Extensible open-source zero-footprint web viewer for cancer imaging research and clinical trials. Cancer Research 77, e119–e122. doi:10.1158/0008-5472.CAN-17-0334.

Medical Imaging and Applications

Master Thesis, June 2019



Weakly Supervised Multi-Organ Multi-Disease Classification using CT

Fakrul Islam Tushar, Joseph Lo, PhD

Department of Radiology, Duke University School of Medicine,Durham,NC.

Abstract

Purpose: Our goal is to investigate using only case-level labels extracted automatically from radiology reports to construct a multi-organ, multi-disease classifier for CT scans with deep learning method.

Methods: In this study, we used a dataset of 23,956 radiology reports from Duke University Health System. We developed a rule-based model to analyze those radiologist reports, labeling disease by text mining to identify cases with those diseases. Initially we focused on three chest-abdomen-pelvis(CAP) organs: lungs, liver, and kidneys. A DenseVNet segmentation model was trained to navigate the target organ from CAP CT scans. Finally a 3D CNN was developed for multi-disease classifications for each organ.

Results: 3D CNN achieved AUC of 0.89 for binary (normal vs. abnormal) and average AUC 0.77 in multi-class for lung, AUC 0.69 for both binary and multi-class for liver, and AUC 0.62 for multi-class kidney classification. As demonstrated with the lung cases, 3D models outperformed 2D models.

conclusion: As an initial baseline this study shows encouraging results of using weak supervision. Further extension of the thesis can be to experiments with larger dataset with more disease types.

Keywords: Weak-supervision, 3D CNN, CT, Lung , Liver, Kidney, Segmentation, Rule-based Model, Radiology reports.

1. Introduction

Computed tomography (CT) scans is one of the most common radiological screening examinations produces 3D images with more finer details compared to standard Chest X-rays. In the U.S. alone, more than 80 million CT exams are performed each year. This helps in improving the detection capability of chest-abdomen at early stages and hence allows for better treatment options. The voxel values in CT scans represent the radiodensity of the tissues in the Hounsfield scale (HU) which provides high-quality images with high contrast.

With the improvement of modern screening modalities like CT scan, accurate detection becomes a major part of the computer-aided diagnosis (CADx). Automatic classification of diseases using CT generally follows a common pipeline. First, the suspicious candidates are selected from the CT scans. After that, some system includes a segmentation step to separates the region of interest (ROI) from the background in order to remove unnecessary information. Handcrafted features are then extracted from the ROI followed by a classifier, which is trained to estimate the final disease classification.

CNNs are the current standard techniques for image classification. Especially for natural images both pre-trained models and training from scratch models shown surprisingly good accuracy. Although, the majority of the existing studies in the field of medical imaging target one particular disease type which make them less effective in practice. In practice, the radiologist can observe multiple finding and they often relate mutually. one of the reasons for this limitation is the availability of well-annotated medical data. It becomes truer in case of CT scans where one CT volume could contain 1000 or more slices. Typically, the publicly available datasets either focused on one organ or on specific disease type such as The Lung Image Database Consortium (LIDC) (Armato III et al., 2011), 2017 Kaggle Bowl challenge

(kag, 2017) and, ChestX-ray8 (Wang et al., 2017). Gibson et al. (2018) used ImageNet state of art models (VGG16, Xception, InceptionsV3, Resnet50) as feature extractor and used Naive Bayes, MultiLayer Perceptron (MLP), Support Vector Machine (SVM), Near Neighbors (KNN) and Random Forest (RF) as lung nodule classifier. Nibali et al. (2017) proposed the use of Resnet (He et al., 2015) architecture and explored effect of curriculum learning, transfer learning and varying network depth on malignancy classification . Christodoulidis et al. (2017) pre-trained their network on publicly available texture datasets which then fine-tuned on lung tissue data. Afterward ensemble of the networks were performed to fuse their knowledge. Gao et al. (2018) proposed ILD imaging patterns classification uses the entire image as a holistic input of size 224x244 pixels.

The segmentation of organs is often an important first step in computer-aided detection pipelines and allows quantitative analysis of clinical processes such as diagnostic, treatment planning and treatment delivery (Litjens et al., 2017). Segmentation is the task of classifying each of the voxels to a region. Multi-organ chest-abdominal CT segmentation is a challenging task due to the anatomical variability in organ shape, appearance, and soft tissue deformation (Fries et al., 2017). Manual segmentation of 3D chest-abdominal images is labor-intensive, expensive and not suitable for most clinical workflows. Traditional multi-organ segmentation techniques can be categorized into statistical models (SM) (Cerrolaza et al., 2015) , multi-atlas label fusion (Xu et al., 2015), (Tong et al., 2015) and registration-free methods. Statistical models (SM) and multi-atlas label fusion methods highly dependent on registration for segmentation which is challenging due to high inter-subject variability.Registration-free methods typically train a voxel-wise classifier on unregistered images and mostly relied on hand-crafted organ-specific image features (Badura and Wieclawek, 2016).CT scans are volumetric in nature and 3D context is known to be more helpful to differentiate between disease patterns. However, existing most of the CT studies are in 2D.

This study will attempt to do multi organ, multi-disease classification, and will use chest abdomen pelvis (CAP) CT because it covers wide variety of organs and disease over most of the torso. Specifically, we will focus on three organs lungs, liver and kidneys, which were chosen because they represent large organs with very different anatomical appearance, location, and range of common diseases.

The lung is mostly consisted air, the liver is a dark reddish-brown organ located in the upper right-hand portion of the abdominal cavity, and the kidneys lie on either side of the spine in the retroperitoneal space between the parietal peritoneum and the posterior abdominal wall, well protected by muscle, fat, and ribs. Considering the immense variations of the anatomical structures of chest-abdomen even experienced radiologists can fail to correctly identify the diseases.

Worldwide a substantial proportion of people suffers from chest-abdominal diseases. An estimated 0.15 million deaths in the U.S. from lung cancer comprising approximately 25% of all cancer deaths (ALA, 2019). The liver disease accounts for approximately 2 million deaths per year worldwide, which account for 3.5% of all deaths worldwide (Asrani et al., 2019). In 2016, nearly 0.13 million people in the U.S. treatment for end-stage kidney disease (ESKD), and more than 0.73 million (2 in every 1,000 people) were on dialysis or were living with a kidney transplant (CFD, 2019). According to Centers for Disease Control and Prevention in (CFD, 2019), estimated 15% of adults in the U.S more than 20 million people are thought to have kidney diseases to some degree and the condition is most common among adults older than 60. Early screening and detection have the potential to dramatically improve the survival rates by finding the disease at an earlier stage when it is more likely to be curable. Over 12,000 lung cancer deaths could be prevented if early detection was possible for high-risk patients (Cheung et al., 2018). For that reason developing robust automatic distinction systems is a critical step.

2. State of the art

Computer-aided detection/diagnosis (CADe/CADx) using deep learning in CT images is been an active area for many years. In spite of that, very few studies have been conducted focused on developing multiple organs and multiple disease prediction systems with weak supervision using machine learning. Commonly, a lot of studies can be found targeted one particular disease type such as lung nodules, lung pneumonia, and liver lesion.

Wang et al. (2017) used ChestX-ray8 dataset and textmined eight disease image labels from the radiological reports using natural langiage preprocessing. Then classified the X-ray images using weakly supervised multi-label image classification and disease localization framework. Tang et al. (2018) used attention-guided curriculum learning (AGCL) for joint thoracic disease classification and weakly supervised localization using chect X-rays. Image-level disease labels and severity level information of a subset of data is been used . This severity level information contributed to curricular learning. Furthermore, they used the CNN generated disease heatmaps (visual attention) of confident seed images to guide the CNN in an iterative training process.

Wang et al. (2017) and Tang et al. (2018) both use X-rays 2D images and shown promising results trough weak-supervision. CT being an volumetic data make it more harder to deal in the field of weak-supervision compared to X-rays. Due to computational expense it's hard to feed the complete CT volume in deep models, and using slice-level is challenging as particular disease

could be belongs to only 10-15 slices out of whole volume.

Yan et al. (2019) proposed a lesion annotation network (LesaNet) based on a multilabel CNN to learn the label from the radiology reports associated with the lesion images. Peng et al. (2019) used a multi-head self-attention mechanism to handle the long-distance information in the sentence, and to jointly correlate different portions of sentence representation subspaces in parallel extract various clinical attributes from radiology reports. Segmentation is the most common field of applying deep learning to medical imaging (Litjens et al., 2017). Segmentation of volumetric images face particular challenges due to the need to process large volumetric images under memory constraints and multi-organ segmentation poses additional challenges as more information must be propagated through the network.

Despite these challenges, deep learning showed promising results in multi-organ chest-abdominal CT segmentation. Zhou et al. (2016) segmented 19 abdominal organs on 2D slices in multiple views and showed combined results using majority-voting label fusion. Roth et al. (2017) used 3D Unet (Ronneberger et al., 2015a) to segment 7 organs using a two-stage hierarchical pipeline. Hu et al. (2017) segmented 4 organs using a 3D FCN to generate organ probability maps as features for a level-set-based segmentation. Larsson et al. (2017) used MALF to identify an ROI for each organ and a 3D FCN with hand-tuned input features to complete the segmentation. In our work, to support the target and navigation of chest-abdominal organs DenseVnet (Gibson et al., 2018) was adopted as segmentation framework.

Ke Yan (2018) mined the bookmarks by the radiologists from the PACS and developed a dataset with 32,735 lesions in 32,120 CT slices of 4,427 unique patients. Using this dataset they proposed a lesion detector based on a regional convolutional neural network (RCNN). Yan et al. (2018) proposed 3D context enhanced region-based CNNs (3DCE) to incorporate 3D context into 2D regional CNNs. Getting the slice level bookmarks from the Deeplesion dataset Ke Yan (2018) they generate feature maps separately from multiple neighboring slice which were then aggregated for final prediction. one of the big limitation of these two studies is that these networks do not predict the type of each detected lesion. This type of detection systems are useful to find out the suspicious regions but not able to find specific disease types.

Attempt to address these challenges, this study presents a weakly supervised multi-organ multi-disease 3D classification workflow using chest-abdominal CT. The contribution of this work is three-fold:

a) We proposed a rule-based model that can extract high-accuracy case-level labels from the unstructured CT reports. We hypothesize this as process of weak-supervision. The classification model will learn disease patterns from these case-level labels.

b) For segmentation, we trained DenseVNet with normal chest-abdominal CT volumes, afterward Fine-tuned the CNN with diseased CT volumes aiming to transfer the organ pattern generalization capability. This segmentation step supports the navigation of the targeted organ in the classification task.

c) We developed weakly supervised 3D multi-disease classifiers for lung, liver and kidneys. Our weak supervision is based only on the radiology reports using the rule-based model, without requiring human experts to ever look at any images.

3. Material and methods

The proposed weakly supervised multi-organ miltidisease classification consists of three major step: (a) Disease label mining, (b)Segmentation of chest-abdomen CTs and (3)finally classification with weak-supervision using CT sans.

3.1. Disease Label Mining

Machine learning algorithms are widely used in many kinds of computer vision tasks and have achieved high performances. Despite the recent advances, their applications have been mostly limited to well-annotated large image datasets like ImageNet, MSCOCO, etc. In the medical domain, however, there are no similar large-scale labeled image datasets available and, it is hard and expensive to label medical images directly. This is particularly true for cross-sectional imaging modalities such as CT, where a single scan may contain 1,000 or more slice images, creating not only a clinical challenge for radiologist interpretation but also a research roadblock because of the impracticality of manually labeling disease in large numbers of slices or volumes. Therefore, we are exploring an alternative approach to provide case-level labels of medical images based on the existing radiology reports. A radiology report provides a diagnostic imaging referral and used documentation purposes. Although there exist some guidelines for reporting, reports mostly contain free text, often organized in a few standard sections. In this study, we developed a rule-based model that can quickly identify certain types of abnormalities within CT reports with high accuracy. We code reports as normal if they do not contain any diseases, and as diseased if the radiologist reported the presence of that disease.

3.1.1. Input Data

From CT scans conducted at Duke University Health System from January to April in 2017, we downloaded

CT abdomen and pelvis without IV contrast, X/X/XXXX

Comparison: CT abdomen pelvis without contrast XXX XXX, XXXX Indication: T86.19 Other complication of kidney transplant, T83.89XA Other specified complication of genitourinary prosthetic devices, implants and grafts, initial encounter (HCC), retained ureteral stent.

Technique: CT imaging from the level of the kidneys to the pelvis was performed without intravenous or oral contrast. The patient was scanned in the prone position. Coronal reformatted images were generated and reviewed to assist with anatomic localization and lesion detection.

Findings: Evaluation of the solid organs in the abdomen and pelvis is limited by the lack of IV contrast. Heart is mildly enlarged. No pericardial effusion. Coronary atherosclerosis. Visualized portions of the lung bases are clear. No pleural effusion. Liver contour is smooth. Gallbladder is mildly distended with biliary sludge and calcifications along the inferior margin. Spleen is normal in size. Adrenal glands are unremarkable. Multiple cystic lesions are seen in the bilateral atrophic kidneys. Pancreas is atrophic. There is scattered areas of calcification within the atrophic pancreas, possibly representing calcific chronic pancreatitis. Surgical resection of the right ureter is noted. Postsurgical stranding in the anterior abdominal wall. Mild anasarca. Postsurgical changes are noted in the stomach consistent with prior roux en y gastric bypass. Small bowel is nondilated. Colon is noninflamed. Transplant kidney is noted in the right lower quadrant. There are multiple renal stones noted within the transplant kidney. The ureter is diffusely thick-walled and dilated with evidence of stones within the dilated ureter. A stent is noted in the proximal portion of the dilated ureter with tip in the bladder. Calcified uterine fibroids. Small amount of free fluid is noted in the pelvis. No free intraperitoneal air. Aorta is diffusely atherosclerotic. The aorta is nonaneurysmal. Fusion of the bilateral SI joints. No acute fracture. Multilevel lumbar spondylosis.

Impression: 1. Moderate transplant kidney hydronephrosis with multiple nonobstructive calculi noted within it, similar to prior exam. The transplant ureter is diffusely thick-walled, although unchanged from prior exam. A transplant ureteral stent is noted, which has mildly migrated, although is still within the renal pelvis. 2. Stable appearance to bilateral atrophic kidneys with hyperattenuating cystic lesions. Electronically Reviewed by: XXX, MD Electronically Reviewed on: X/X/XXXX X:XX I have reviewed the images and concur with the above findings. Electronically Signed by: XXX, MD Electronically Signed on: X/X/XXXX X:XX

Figure 1: Example language from radiology reports for CT scans.

23,956 radiology reports. Three organ systems were selected for this study: lungs, liver/gallbladder, and kidneys. Example of an unstructured CT report we used is shown in figure 1. Each report is free text, but typically composed of parts for indication, technique, findings, and impression. Only the findings part is used for our rule-based model. Sentences pertaining to lung. liver, and kidney have been highlighted in yellow, blue and green respectively.

3.1.2. Selecting keywords

In order to select the keywords for each organ, we computed term frequency-inverse document frequency (TFIDF) for all words in each organ, which can reflect how important a word is to a corpus. To illustrate the TF-IDF values, we first identified sentences for each organ based on anatomical keywords (such as for lungs: lung, pulmonary, airway, airspace, etc.), then used each set of anatomical keywords to filter the reports and made a word cloud for each organ like shown in Figure 2. In this way, we could find out which word is more important for our rule-based system. The clinical experts provided guidance on whether the terms are negative (e.g., normal, no evidence of, unremarkable), positive (e.g., abnormal, status post, *itis), or do not matter.

3.1.3. Rule-based system

With the keywords, we designed a rule-based system to classify selected, common diseases and normal reports. Figure 3 shows the rule-based model's workflow and the logic. From the unstructured radiology reports we first extract the findings section. Afterward the text was converted to lowercase and sentence tokenization was performed. As a final step the logic operation was applied to code reports as normal if they do not contain any diseases, and as diseased if the radiologist reported the presence of that disease. For the lungs, we focused on normal, atelectasis, nodules, pneumonia, and pulmonary edema. The liver and gallbladder were considered as one organ group, and we focused on normal, masses, and stones. For the kidneys, we focused on normal, atrophy, stone, and masses. The rules were intentionally very specific to maximize positive predictive value. Each combination of organ and disease resulted in many hundreds to thousands of reports. The number of cases was 16304 for lungs, 9286 for liver/gallbladder, and 9700 for kidneys, for a total of 35290 rule-based labeled cases. Clinical experts checked each category and the model would be considered acceptable if at least 50 consecutive reports were labeled with 95% accuracy. Almost all categories were 100% accurate.

3.2. Segmentation

In this study, we aim to use the segmentation model as a prepossessing step or more precisely to support the navigation and localization of the chest-abdomen organs for the classification of abnormalities in CT volume. Figure 4 shows an overall segmentation pipeline used in this study including data splitting, pre-processing and evaluation.

3.2.1. Segmentation data

The CT volumes used in this study for the training of the segmentation models are called 4D extended cardiac-torso (XCAT) phantom, developed by a team from Duke University Medical Center for multimodality imaging research based on Chest-Abdomen-Pelvis (CAP) CTs studies (Segars et al., 2013). There are 50 CT volumes and almost all of them have intravenous contrast enhancement. Selected organs and structures were manually labeled and have been continuously refined over the years. Currently, the dataset contains 29 different classes of structures available, including: sternum, ribs, spine, pelvis, scapula, clavicles, femur, heart,

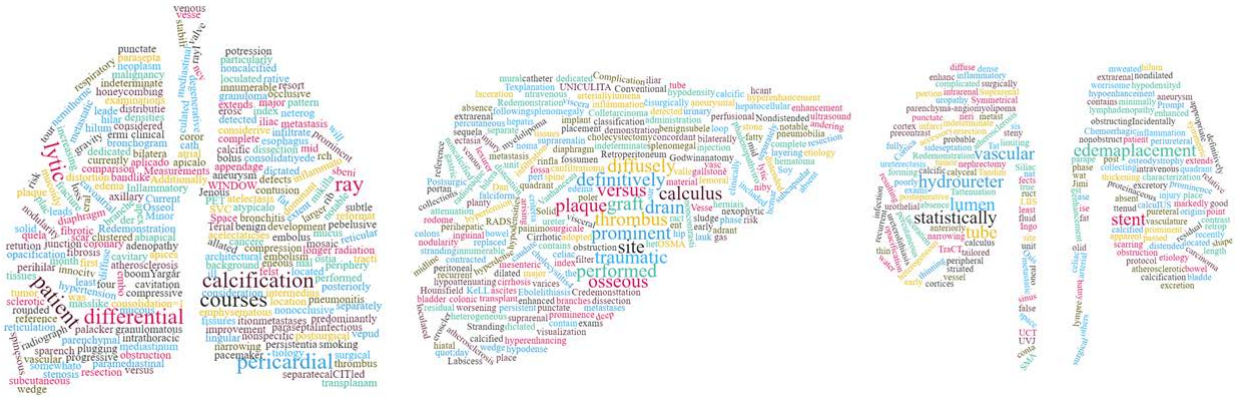


Figure 2: Word clouds generated from report sentences for (left to right) lung, liver/gallbladder, and kidneys, based on TFIDF.

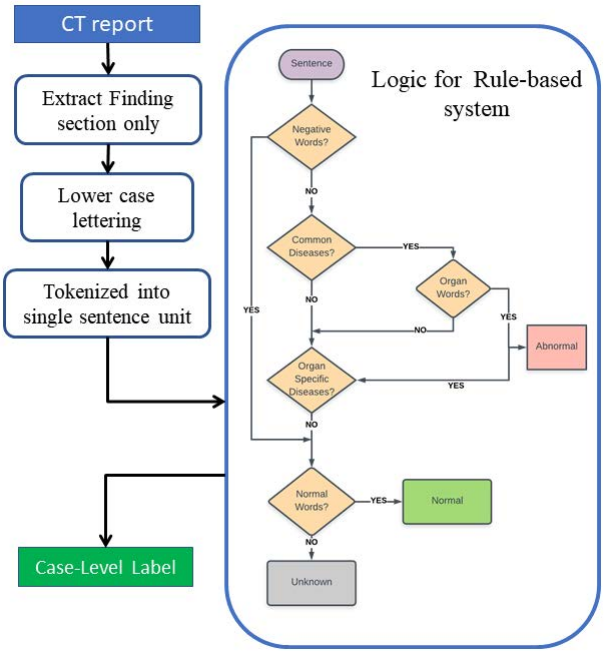


Figure 3: Rule-based model

lungs, liver, gall bladder, stomach, spleen, thyroid, kidneys, pancreas, large intestines, skin, arm, bladder, abdomen muscle and background.

In this study we particularly focused on lungs, liver and kidneys. Each CT slice and the corresponding segmentation map (label slice) are of size 512×512 pixels. The resolution of each CT volume is approximately 0.7×0.7 mm, where the thickness of each slice varied between 0.6-5.0 mm. The XCAT dataset (50 patients) was randomly split into 44 CT volumes for training, and 6 CT volumes for validation. Moreover, we incorporated 50 more recent patients with high-resolution CTs with 0.7 mm slice thickness and manually labeled kidney segmentation masks for the purpose of quantitatively evaluating the performance of our trained segmentation model.

3.2.2. Pre-processing

Based on population statistics, we selected a range of intensity values [-1000, 800] for all CT volumes, followed by a linear transformation, to obtain a normalized and well-covered spectrum of CT intensity values across CTs in the dataset. For the low-resolution (XCAT) data, each voxel in the original CT volume represented an intensity covering approximately a $0.7 \times 0.7 \times 5$ mm volume, which varied from patient to patient. We re-sample the resolution across studies via B-spline interpolation into voxels of size $2 \times 2 \times 2$ mm. In order to keep the segmentation labels consistent, we applied a similar transformation on the labels using nearest neighbor interpolation. We saved the transformation metrics in case we need to go back to the original spacing for some interpretation.

3.2.3. Segmentation Model

In the first stage of our segmentation study, we implemented three well-known CNN architectures: 3D U-net (Ronneberger et al., 2015b), FCN (Long et al., 2014) and DenseVNet (Gibson et al., 2018). Based on the evaluation of the performance on the XCAT validation set, DenseVNet was selected as the segmentation model in our workflow.

The DenseVNet segmentation network uses fully convolutional neural network, with high-resolution activation maps through memory-efficient dropout and feature reuse. Gibson et al. (2018) used a patch size of 144^3 as an input of the network. First, 72^3 feature maps are computed using a strided convolution. Afterward, a cascade of dense feature stacks and strided convolutions generate activation maps at three resolutions. The downsampling subnetwork is a sequence of three dense feature stacks connected by downsampling strided convolutions (Gibson et al., 2018). A convolution unit is applied at each resolution reducing the number of features by half. After bilinear upsampling, the feature maps are concatenated and a final convolution generates the likelihood logits. Memory efficiencies of dense fea-

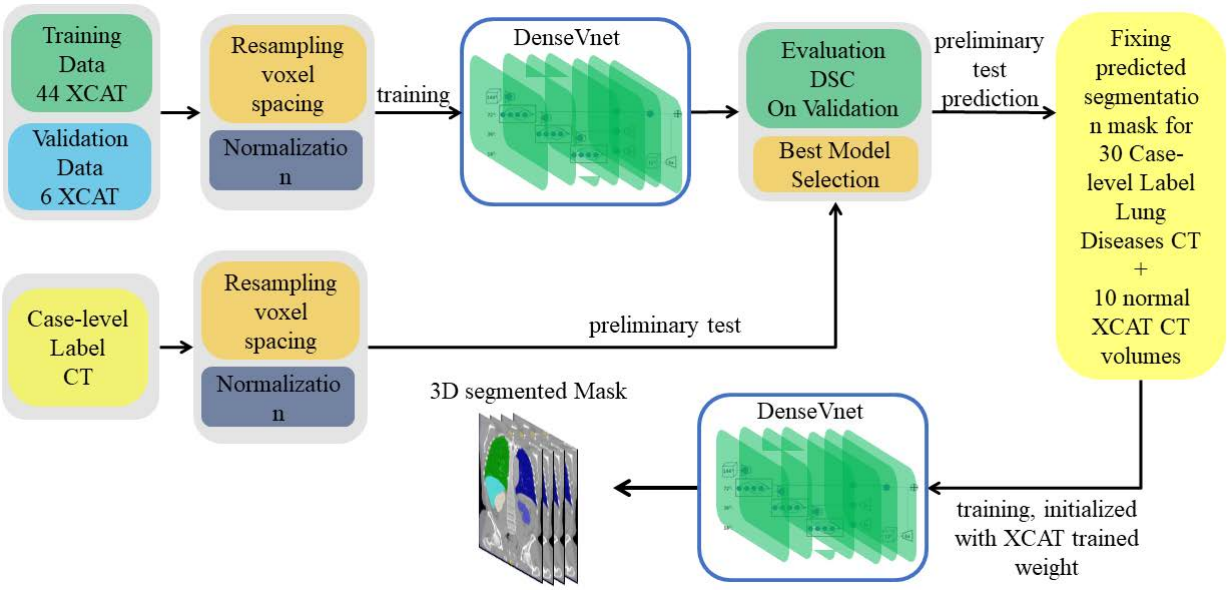


Figure 4: Complete organ segmentation pipeline

ture stacks and batch-wise spatial dropout enable deeper networks at higher resolutions, which is advantageous for segmentation of smaller structures.

3.2.4. Training segmentation model with diseased CT volumes

Further to achieve good results for the Duke diseased CT cases, the model was fine-tuned with 30 abnormal and 10 normal Cheat-abdominal CT scans. As mentioned earlier in Section 3.1 in this study for the classification we are using only case-level labels, so no segmentation annotation mask was available for these disease cases.

The XCAT trained model was used to generate a preliminary mask for 30 randomly selected, diseased lung cases (10 for edema,10 atelectasis, 7 for pneumonia, and 3 for nodule). The initial predicted segmentation mask was then manually corrected only for lungs. Afterward combining this 30 annotated cases with 10 normal XCAT training volume we trained the model from scratch using the pre-trained weights from XCAT trained model.

3.3. Weakly Supervised classification

3.3.1. Classification data and Pre-procesing

From January to April of 2017, there were approximately 5,000 chest CT scans at our institution, Duke University Health System. After applying our rule-based model, we identified many hundreds to over a thousand cases for each category. Table1 shown the list of data used in our classification tasks.These diseases were selected based on their highly specific keywords, as well as to represent a variety of disease sizes and appearances. We grouped the same disease types to one class of disease. With IRB approval, these cases were downloaded and de-identified. Same pre-processing scheme mentioned in section 3.2.2 is been applied.

3.3.2. 2D CNNs for weakly supervised classification

We used Resnet50 (He et al., 2015) as the model because of its advantages reducing the effect of the vanishing gradient problem through residual blocks. We removed its last fully connected layer, and then added three fully connected layers with 2048,2048 and 5 neurons, respectively. These newly added fully connected layers were initialized using Xaiver algorithm and the rest of the model was initialized using ImageNet weights. Softmax activation function was used for the last layer.

3.3.3. 2D Training Phase

With the 2D model we focused only on the lungs as a single organ, and characterized abnormality in terms of four lung diseases with high prevalence: atelectasis, edema, nodule, and pneumonia. Classification is performed considering these as four diseases as separate classes, combined with normal as a fifth class. For each case, we chose 10 slices spaced regularly in the superior to inferior extent of the volume. The network was fine-tuned end-to-end using Adam with standard parameters. We trained the model for 50 epochs using batch-size of 32 and an initial learning rate of 0.001. We picked the model with the lowest validation loss.

The 4-fold cross-validation performance was measured by the Receiver Operating Characteristic (ROC) area under the curve (AUC). We present the slice-level performance, where each slice is treated separately, i.e., the patient is essentially being diagnosed using only a

Table 1: Dataset Distribution used in CT classification for lungs, liver/gallbladder and kidneys. + sign denotes the grouping of same disease type to single class.

Lungs		Liver/Gallbladder		Kidneys	
Volume	**# of Volumes**	**Volume**	**# of Volumes**	**Volume**	**# of Volune**
Normal	137	Normal	181	Normal	68
Edena	214	Dilation	54	Cyst	61
Atelectasis	201	Lesion+Mass+Nodule	102	Lesion+ Mass + Tumor	52
Pneumonia	101	Calci+Gallstone+Stone	70	Calcifi+Calcul+Stone	66
Nodule	168			Atroph+ Atrophy	58
Total	821	Total	407	Total	305

single slice. We also presented a simple ensemble approach to get a patient-level performance, where we average the probability for all 10 slices as the prediction score for each patient.

3.3.4. 3D CNNs for weakly supervised classification

The baseline 3D CNN used in this study was inspired by Resnet (He et al., 2015). Figure 5 shown the proposed 3D CNN architecture. One initial convolution was performed on input volumes, afterward, features were learned in three resolution scales using 3 R-blocks unit in each resolution. An R-block consists of Batch-normalization, reLu activation, and 3D convolution. Long et al. (2014) showed that deeper network have greater discriminative power due to the additional non-linearities and better quality of local optima. However, convolutions with 3D kernels are computationally expensive and 3D architectures have a larger number of trainable parameters. We used 3^3 kernels that are faster to convolve with and contain fewer weights. Batch-normalization allows normalization of the feature map activation at every optimization step. After each resolution, the features were reduced half by max-pooling and the number of filters is doubled. After the 3rd resolution last R-block there were $16^3 \times 128$ features, which passed through batch-normalization, relu followed by a global max-pooling and finally softmax classification layer for the final prediction.

3.3.5. 3D Training Phase

As with the 2D classification experiments, it would be desirable to do a 4-fold cross-validation to reduce the variance due to data sampling. However, 3D models are extremely time consuming to develop, with run times of 1 day per fold with typical GPU cards such as the GeForce GTX 1080 (Nvidia Corporation, Santa Clara CA). Due to time constraints, for all 3D models, results shown are the average of 2 of the 4-fold cross-validation samplings.

We applied our fine-tuned segmentation model on the classification dataset to get the chest-abdomen organ segmenation masks. These segmentation masks guide the 3D model to extract desired organ patches on the fly. Due to the computational expense it's not practical to feed the whole CT volume into 3D CNN. We extracted 2 patches of size $128 \times 128 \times 128$ from each volume using the segmented mask. While extracting the patches the targeted organ labels got the highest preferences. Adam was used as a optimizer to optimize the weights. Cross-entropy error was used as the loss function and the weights were updated using batch of 2 training samples for every iteration. Initialization of the weights was done by uniform distribution. Training was continued for 50000 iteration and only the weights for the lower loss on the validation set was saved.

To start with the 3D experiments, We first explore the performance of our weakly supervised model on lungs CT scans. As mentioned in Table 1 for the lungs we have five classes: normal, edema, atelectasis, pneumonia, and nodule. We first combined all the lung disease classes (edema, atelectasis, pneumonia, and nodule) into a single class as abnormal, and classified normal vs abnormal. Afterward we separated all the 5 classes and trained the model for a 5 class classification.

For the liver we focused on three diseased classes: dilation, masses (lesion, mass, nodule), stones (calci*, gallstone, stone), and normal class. Similarly in case of liver we first combine all the liver disease classes: dilation, masses, and stones to a single class as abnormal and classify normal vs abnormal. Afterward we separate all the 4 classes and train the model for a 4 class classification.

For the kidneys we focused on four diseased classes (cyst, masses, stones, and atrophy), and the normal class. Due to time constrain we couldn't experiment with the normal vs abnormal. We separate all the 5 classes and train the model for a 5 class classification.

The model is implemented using python tensorflow framework.

4. Results

4.1. Segmentation Results

The XCAT dataset (50 patients) was randomly split into training (44 patients), validation (6 patients), as mentioned earlier in section 3.2.1. For the quantitative comparison between the different segmentation models dice similarity coefficient (DSC) was used as an evaluation metric. The average validation DSC values for

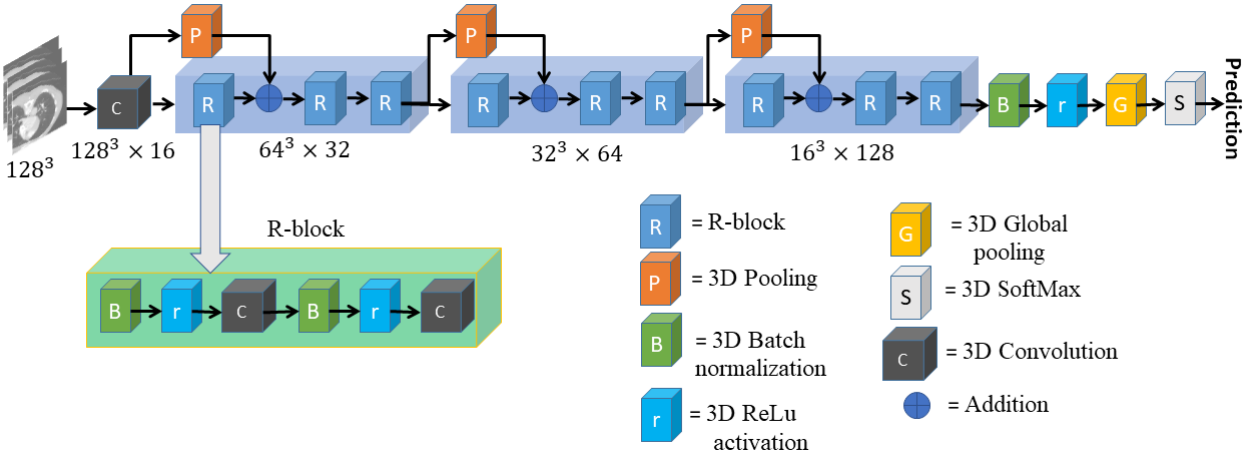


Figure 5: Our baseline 3D CNN With 3 R-Blocks in each resolution. Number of FMs and their size depicted as (Size Number)

each organ are reported in Table 2. Figure 6 shown a randomly selected CT scan from validation set,and corresponding ground-truth and prediction.

Figure 7 shows an example of lung edema and lung atelectasis cases and the segmented labels by the model before and after fine-tuning. Since there is no ground-truth segmentation for the Duke data, the segmentation labels were qualitatively evaluated by Duke radiology experts. Results for all 50 Kidney test cases, including 7 without contrast, are shown in Figure 8.

4.2. Classification Results

4.2.1. Lung Diseases Classification

Figure 9 and figure 10 shown the slice-level and patient-level performance of the 2D model. The patient-level prediction provides better overall performance than slice-level prediction for all classes.

For the 3D classification, Figure 11 shows the binary results for the lungs, i.e., normal vs abnormal classification. ROC curves are averaged across 2 of the 4-fold cross-validation samplings.

Figure 12 show the ROC curves for multi-class classification (normal, edema, atelectasis, penumonia, and nodule) using 3D CNN.

4.2.2. Liver Diseases Classification

In Case of liver 3D experiments, we first combine all the Liver disease classes (dilation, mass, and stone) to a single class as abnormal and classify normal vs abnormal. Figure 13 shown the liver normal normal vs abnormal averaged across 2 of the 4-fold cross-validation samplings

Figure 13 shown the multi-class classification (normal, dilation, mass, and stone) results for the liver. ROC curves are averaged across 2 of the 4-fold cross-validation samplings.

4.2.3. Kidneys Diseases Classification

Figure 12 shown the multi-class classification (normal, cyst, mass, stone, and atrophy) results for the Kidney. ROC curves are 1 of the 4-fold cross-validation samplings.

5. Discussion

In this study, a weakly supervised 3D classification workflow was proposed for the purpose of classifying multiple diseases of lung, liver and kidneys using CT images. Unlike conventional methods which require well-annotated data and handcrafted features, the proposed system adopts a rule-based model to analyze radiology reports to provide case-level labels. These provide a form of weak supervision, because each label (e.g., lung nodule) applies to all 2D or 3D patches within a case, whereas the disease may be present only in some or none of those patches. By providing sufficient numbers of cases, the model can learn disease patterns from these noisy, case-level labels.

5.1. Segmentation

As we are working with chest-abdominal organs, to support the navigation and classification of specific organs, we trained a deep segmentation model. Based on the validation performance, the DenseVNet outperformed both 3D FCN and 3D Unet, with an especially high margin for the kidneys. Over 0.9 DSC for lungs and liver and over 0.8 for kidneys predictive segmentation performance was achieved on the XCAT validation dataset. Compared to the XCAT ground truth established by non-medical research assistants, there were actually many instances where our segmentation model provided more accurate segmentation. This demonstrates the potential improvements from automated algorithms trained with sufficient data.

Table 2: Average Dice coefficient for the evaluation of XCAT validation set . Bold Font Highlights The Highest DSC.

	Right lung	Left lung	Liver	Right kidney	Left kidney
	Dice coefficient (%)				
3D FCN	0.93 0.06	**0.95 0.01**	0.91 0.03	0.84 0.08	0.78 0.07
3D Unet	0.94 0.06	**0.95 0.01**	0.92 0.03	0.86 0.07	0.80 0.07
DenseVnet	**0.95 0.02**	**0.95 0.01**	**0.92 0.01**	**0.89 0.03**	**0.89 0.02**

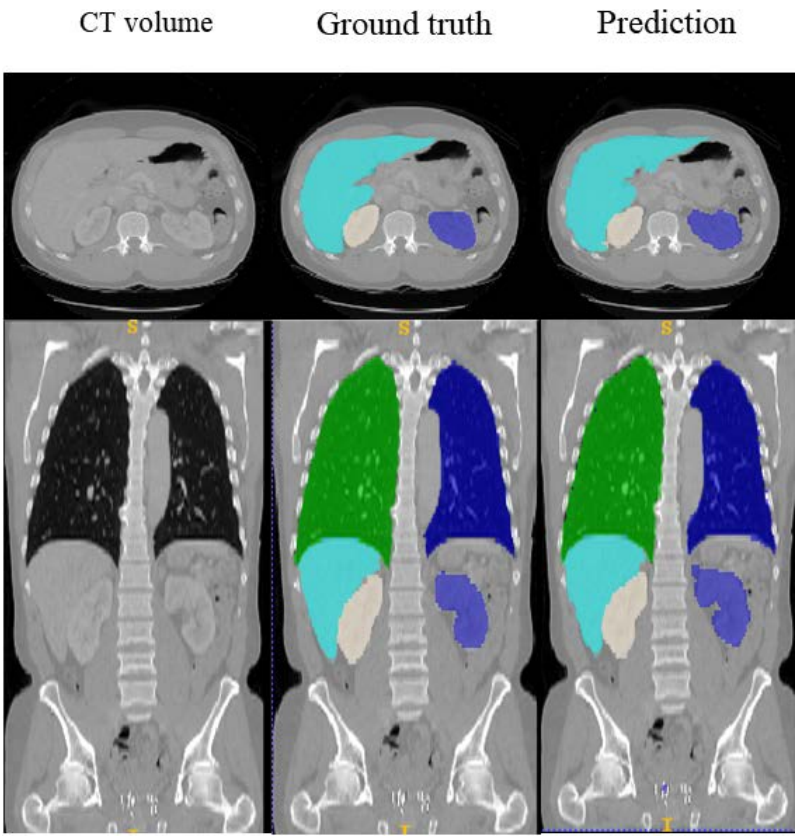


Figure 6: An example of the segmentation.

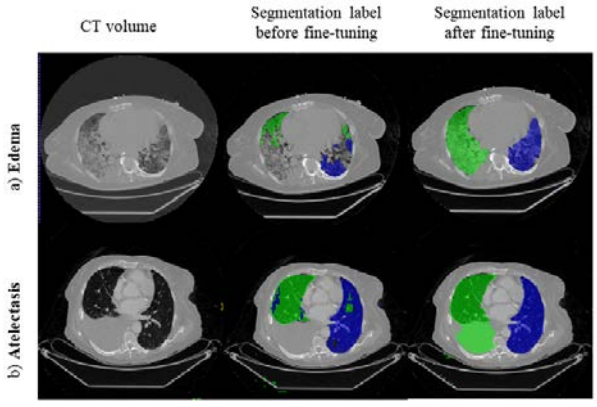


Figure 7: Qualitative results on duke diseased data without ground-truth segmentation labels. Top:Lung Edema,right lung (green), left lung (blue). Buttom:Lung atelectasis CT volume,,right lung (green),left lung (blue).

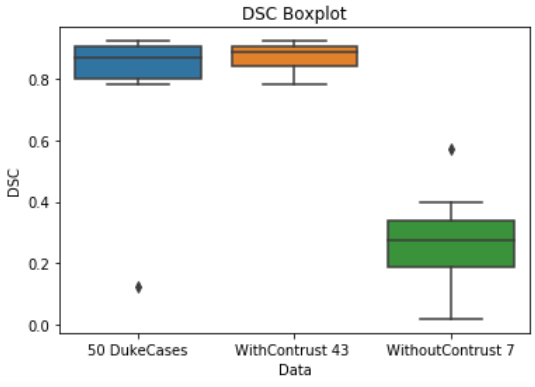


Figure 8: Segmentation results for kidneys on highresolution Duke data.

Although we achieved a high segmentation performance over the validation set of XCAT, the trained model failed to generalize to new cases with diseased

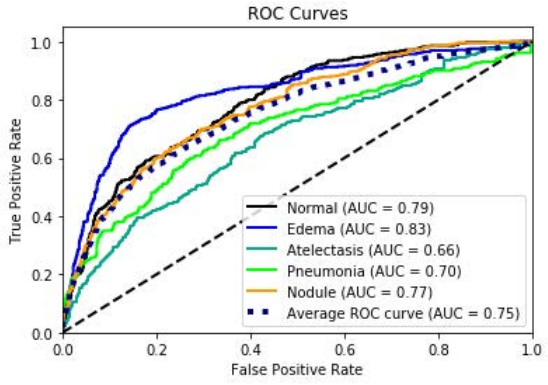


Figure 9: Slice-based ROC curves for 2D multi-class lung classifier (normal, edema, atelectasis, penumonia, and nodule)

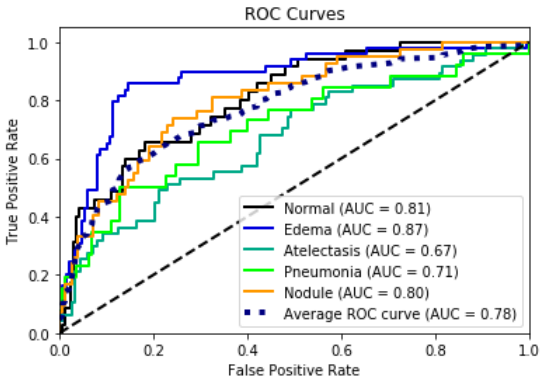


Figure 10: Patient-based ROC curves for 2D multi-class lung classifier (normal, edema, atelectasis, penumonia, and nodule)

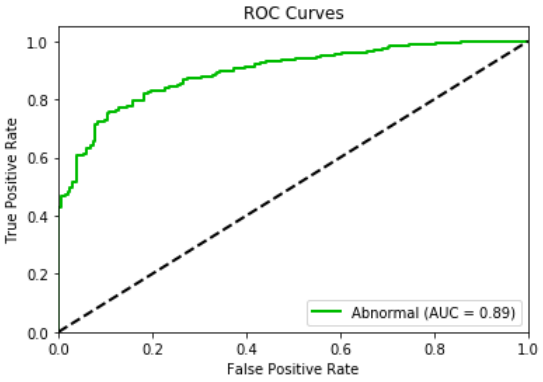


Figure 11: ROC curve for 3D binary lung classifier: normal vs abnormal (combination of edema, atelectasis, penumonia, and nodule)

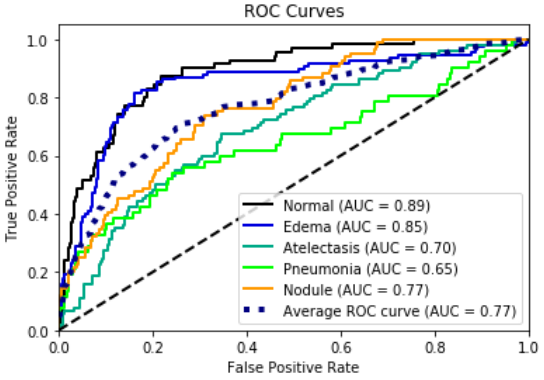


Figure 12: ROC curves of 3D multi-class lung classifier (normal, edema, atelectasis, penumonia, and nodule)

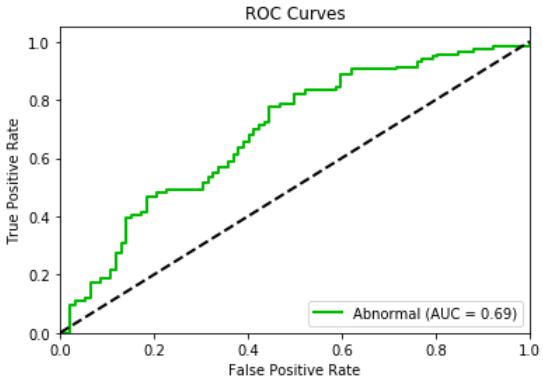


Figure 13: ROC curve for 3D binary Liver classifier: normal vs abnormal (combination of dilation, mass, and stone)

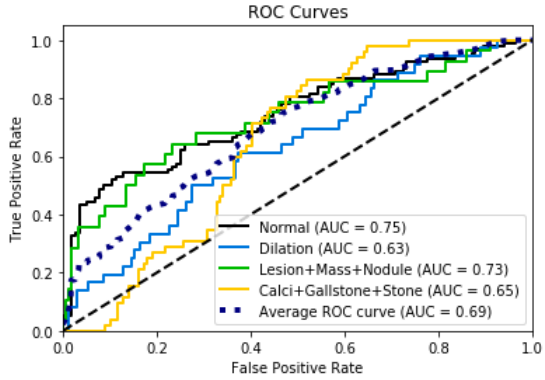


Figure 14: ROC curves of 3D multi-class liver classifier (normal, dilation, mass, and stone)

lungs. This downfall in performance can be explained by the fact that all XCAT CT volumes (training and validation set) belong to the normal class and no pathological abnormalities are present. Normally lung anatomy consists mostly of air, but this can change with certain diseases, which is why we selected lung diseases such as edema which fill the lungs with fluid instead of air, or atelectasis which can collapse the entire lung or area (lobe) of the lung. So during training, the segmentation model have not learned these types of abnormalities to handle during testing. However, our fine-tuning process using manually corrected disease cases provided dramatic qualitative improvement in segmentation. Moreover, this semi-automated process is relatively quick and easy, and demonstrates a potential workflow to improve performance by enriching the training with difficult cases.

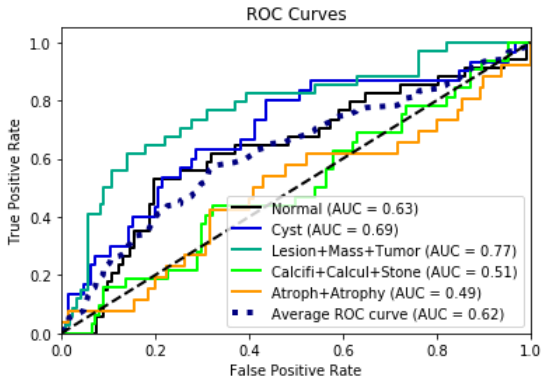


Figure 15: ROC curves of 3D multi-class kidney classifier (normal, cyst, mass, stone, and atrophy)

Using the lung cases, we demonstrated a technique to pool together the predictive results from individual CT slice images, producing a patient-level prediction that provided better overall performance across all five classes. This improvement in performance is expected, as during slice-level prediction we labeled the slices based on the radiology report and all slices shared the same label. So the label was weak because often the disease will not appear in all slices. When we averaged the predictions for all slices of each patient, the high predicted probabilities helped to improve performance. Pooling by the maximum instead of the average may further enhance performance, especially for focal disease such as lung nodules.

Moving from 2D CNN to 3D CNN allowed us to take advantage of the volumetric nature of CT, as AUC was improved for 3 out of the 4 lung disease classes. The only exception was for pneumonia, which is logical because out of the 5 classes for the lungs, that had the lowest samples, which affected the learning curve of the 3D CNN. This performance improvement is in the tradeoff of more computation power, time and expenses. Although, our main goal is to investigate the effect of weak supervision on multi-disease classification, experimenting with normal vs. abnormal gives us a fair idea of how hard it's for that network to learn discriminate features.

In the case of liver, in multi-class classification, we have a huge imbalance in data where the normal class is almost 50% of the entire liver dataset. This reason supports why the perforce outcome for both the binary and multi-class AUC are exactly the same. As we continue to explore the relative advantages of these two approaches toward classification, it will be important to keep the class prevalence in mind. In fact, one very important thing to notice in both 3D multi-class classification task for lung and liver is that the worst performing class is the class which has the lowest samples.

5.2. Limitations

Since this is our very first attempt towards multi-organ, multi-disease classification, it has a few limitations. Training a deep model requires a considerable amount of data, with the most successful studies in the literature sometimes boasting tens to hundreds of thousands of cases. For our study, each disease class is near 200 CT volumes, with each organ represented by approximately 1000 cases. Although this is better than many other studies, in the perspective of deep learning, this number is likely still insufficient. Downloading these CT volumes from PACS and performing de-identification is a time-demanding task, and also storing these volumes (each 0.5-1.0 GB) requires a lot of resources due to data size. In ongoing work, we are in the process of doubling the data which will likely improve model performance and generalizability.

Compared to 2D CNN, working with the 3D CNNs is computationally very expensive, so it becomes highly nontrivial in case of CT volumes that can be $512\times512\times1000$ to process such large volumetric data. Due to our time constraint, we couldn't analyze the complete performance of the 3D network. Our future experiments will investigate important issues such as optimal patch size for training, relationship of network depth with features learned, optimal learning rate, and hyper-parameter tuning.

We also observed that the segmentation performance was much worse for the minority of cases without iodine contrast enhancement. This reflects the distribution of cases in the clinical data. In the long term, it will be advantageous to enrich the training data with less common cases acquired without contrast or with lower dose protocols, or even cases with known image quality artifacts. This will enrich the training and improve the generalization of the system.

6. Conclusions

In this study, we proposed a weakly supervised, multi-organ, multi-disease classification framework that uses a rule-based model to provide case-level labels. A deep segmentation model provided navigation to target organs in chest-abdomen-pelvis CT scans. Finally, we performed the multi-disease classification using 3D CNN. To our best knowledge, this is the first medical imaging study where multiple, diverse organs were targeted to classify multiple, diverse diseases. As an initial baseline, this study shows encouraging results of using weak supervision and opens up a new, exciting, and unexplored field. Further extension of the thesis can include experiments with a larger dataset and more disease types.

7. Acknowledgments

I would like to acknowledge Rui Huo, Yinhao Ren, Ruixiang Tang, Songyue Han, Geoffrey D. Rubin and Brian Harrawood who are part of Duke Radiology group and were part of weekly meetings. They helped me refine different ideas and also provided access to their GPUs when required.

References

, 2017. Data science bowl 2017. URL: https://www.kaggle.com/c/ data-science-bowl-2017/ overview/engagement-contest.

, 2019. American lung association.lung cancer fact sheet. Available at:https://www.lung.org/ lung-health-and-diseases/lungdisease-lookup/lung-cacer/ resource-library/lung-cancerfact-sheet.html. Accessed: 2019-06-05.

, 2019. Chronic kidney disease in the united states, 2019 .

Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., Kazerooni, E.A., MacMahon, H., van Beek, E.J.R., Yankelevitz, D., Biancardi, A.M., Bland, P.H., Brown, M.S., Engelmann, R.M., Laderach, G.E., Max, D., Pais, R.C., Qing, D.P.Y., Roberts, R.Y., Smith, A.R., Starkey, A., Batra, P., Caligiuri, P., Farooqi, A., Gladish, G.W., Jude, C.M., Munden, R.F., Petkovska, I., Quint, L.E., Schwartz, L.H., Sundaram, B., Dodd, L.E., Fenimore, C., Gur, D., Petrick, N., Freymann, J., Kirby, J., Hughes, B., Vande Casteele, A., Gupte, S., Sallam, M., Heath, M.D., Kuhn, M.H., Dharaiya, E., Burns, R., Fryd, D.S., Salganicoff, M., Anand, V., Shreter, U., Vastagh, S., Croft, B.Y., Clarke, L.P., 2011. The lung image database consortium (lidc) and image database resource initiative (idri): A completed reference database of lung nodules on ct scans. Medical Physics 38, 915–931.

Asrani, S.K., Devarbhavi, H., Eaton, J., Kamath, P.S., 2019. Burden of liver diseases in the world. Journal of Hepatology 70, 151 – 171.

Badura, P., Wieclawek, W., 2016. Calibrating level set approach by granular computing in computed tomography abdominal organs segmentation. Applied Soft Computing 49, 887 – 900.

Cerrolaza, J.J., Reyes, M., Summers, R.M., ngel Gonzlez-Ballester, M., Linguraru, M.G., 2015. Automatic multi-resolution shape modeling of multi-organ structures. Medical Image Analysis 25, 11 – 21.

Cheung, L.C., Katki, H.A., Chaturvedi, A.K., Jemal, A., Berg, C.D., 2018. Preventing Lung Cancer Mortality by Computed Tomography Screening: The Effect of Risk-Based Versus U.S. Preventive Services Task Force Eligibility Criteria, 20052015. Annals of Internal Medicine 168, 229–232.

Christodoulidis, S., Anthimopoulos, M., Ebner, L., Christe, A., Mougiakakou, S., 2017. Multisource transfer learning with convolutional neural networks for lung pattern analysis. IEEE Journal of Biomedical and Health Informatics 21, 76–84.

Fries, J.A., Wu, S., Ratner, A., Ré, C., 2017. Swellshark: A generative model for biomedical named entity recognition without labeled data. CoRR abs/1704.06360.

Gao, M., Bagci, U., Lu, L., Wu, A., Buty, M., Shin, H.C., Roth, H., Papadakis, G.Z., Depeursinge, A., Summers, R.M., Xu, Z., Mollura, D.J., 2018. Holistic classification of ct attenuation patterns for interstitial lung diseases via deep convolutional neural networks. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 6, 1–6.

Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula, S., Gurusamy, K., Davidson, B., Pereira, S.P., Clarkson, M.J., Barratt, D.C., 2018. Automatic multi-organ segmentation on abdominal ct with dense v-networks. IEEE Transactions on Medical Imaging 37, 1822–1834. doi:10.1109/TMI.2018.2806309.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. CoRR abs/1512.03385. arXiv:1512.03385.

Hu, P., Wu, F., Peng, J., Bao, Y., Chen, F., Kong, D., 2017. Automatic abdominal multi-organ segmentation using deep convolutional neural network and time-implicit level sets. International Journal of Computer Assisted Radiology and Surgery 12, 399--411.

Ke Yan, Xiaosong Wang, L.L.R.M.S., 2018. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. Journal of Medical Imaging 5, 1 -- 11 -- 11.

Larsson, M., Zhang, Y., Kahl, F., 2017. Robust abdominal organ segmentation using regional convolutional neural networks, in: Sharma, P., Bianchi, F.M. (Eds.), Image Analysis, Springer International Publishing, Cham. pp. 41--52.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Snchez, C.I., 2017. A survey on deep learning in medical image analysis. Medical Image Analysis 42, 60 -- 88.

Long, J., Shelhamer, E., Darrell, T., 2014. Fully convolutional networks for semantic segmentation. CoRR abs/1411.4038. URL: http://arxiv.org/abs/1411.4038, arXiv:1411.4038.

Nibali, A., He, Z., Wollersheim, D., 2017. Pulmonary nodule classification with deep residual networks. International Journal of Computer Assisted Radiology and Surgery 12, 1799--1808.

Peng, Y., Yan, K., Sandfort, V., Summers, R.M., Lu, Z., 2019. A self-attention based deep learning method for lesion attribute detection from CT reports. CoRR abs/1904.13018. URL: http://arxiv.org/abs/1904.13018.

Ronneberger, O., Fischer, P., Brox, T., 2015a. U-net: Convolutional networks for biomedical image segmentation. CoRR abs/1505.04597.

Ronneberger, O., Fischer, P., Brox, T., 2015b. U-net: Convolutional networks for biomedical image segmentation, in: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention -- MICCAI 2015, Springer International Publishing, Cham. pp. 234--241.

Roth, H.R., Oda, H., Hayashi, Y., Oda, M., Shimizu, N., Fujiwara, M., Misawa, K., Mori, K., 2017. Hierarchical 3d fully convolutional networks for multi-organ segmentation. CoRR abs/1704.06382. URL: http://arxiv.org/abs/1704.06382, arXiv:1704.06382.

Segars, W.P., Bond, J., Frush, J., Hon, S., Eckersley, C., Williams, C.H., Feng, J., Tward, D.J., Ratnanather, J.T., Miller, M.I., Frush, D., Samei, E., 2013. Population of anatomically variable 4d xcat adult phantoms for imaging research and optimization. Medical Physics 40, 043701.

Tang, Y., Wang, X., Harrison, A.P., Lu, L., Xiao, J., Summers, R.M., 2018. Attention-guided curriculum learning for weakly supervised classification and localization of thoracic diseases on chest radiographs, in: Shi, Y., Suk, H.I., Liu, M. (Eds.), Machine Learning in Medical Imaging, Springer International Publishing, Cham. pp. 249--258.

Tong, T., Wolz, R., Wang, Z., Gao, Q., Misawa, K., Fujiwara, M., Mori, K., Hajnal, J.V., Rueckert, D., 2015. Discriminative dictionary learning for abdominal multi-organ segmentation. Medical Image Analysis 23, 92 -- 104.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M., 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. CoRR abs/1705.02315. URL: http://arxiv.org/abs/1705.02315, arXiv:1705.02315.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M., 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3462--3471.

Xu, Z., Burke, R.P., Lee, C.P., Baucom, R.B., Poulose, B.K., Abramson, R.G., Landman, B.A., 2015. Efficient multi-atlas abdominal segmentation on clinically acquired ct with simple context learning. Medical Image Analysis 24, 18 -- 27.

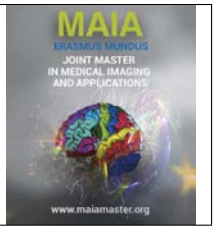
Yan, K., Bagheri, M., Summers, R.M., 2018. 3d context enhanced region-based convolutional neural network for end-to-end lesion detection. CoRR abs/1806.09648. arXiv:1806.09648.

Yan, K., Peng, Y., Sandfort, V., Bagheri, M., Lu, Z., Summers, R.M., 2019. Holistic and comprehensive annotation of clinically significant findings on diverse CT images: Learning from radiology reports and label ontology. CoRR abs/1904.04661. URL: http://arxiv.org/abs/1904.04661, arXiv:1904.04661.

Zhou, X., Ito, T., Takayama, R., Wang, S., Hara, T., Fujita, H., 2016. Three-dimensional ct image segmentation by combining 2d fully convolutional network with 3d majority voting, in: Carneiro, G., Mateus, D., Peter, L., Bradley, A., Tavares, J.M.R.S., Belagiannis, V., Papa, J.P., Nascimento, J.C., Loog, M., Lu, Z., Cardoso, J.S., Cornebise, J. (Eds.), Deep Learning and Data Labeling for Medical Applications, Springer International Publishing, Cham. pp. 111--120.

Medical Imaging and Applications

Master Thesis, June 2019



Radiomics versus Convolutional Neural Networks for Survival Time Prediction and Therapy Response of Metastatic Melanoma in Computed Tomography

Gulnur Semahat Ungan, Supervisors: Adrien Bartoli, Benoit Magnin

Encov Research Group,Institut Pascal, Universite Clermont Auvergne, Clermont Ferrand, France

Abstract

Metastatic melanoma is a malignancy of poor prognosis, although the recent use of immunotherapy has improved global survival. Nevertheless there is a need for prognostic predictors to treatment response in metastatic melanoma. Visual analysis of the CT scan by a radiologist can provide with some predictors (number of lesion, number of metastatic sites, lesion sizes...), but little research has been published on computer aided prediction in metastatic melanoma.
Purpose of this thesis is to compare the diagnostic performance of radiomic features and a convolutional neural network (CNN) for classification of the metastatic melanoma patients in terms of 1 year survival and therapy response to immunotherapy.
Manual segmentation (volumic (3D) and one slice (2D)) was performed on CT images. Radiomic features were extracted from the segmented lesions, then selected by 5 different feature selection algorithms and classification performed using 4 different algorithms. Same segmented lesions were used to feed the Deep Neural Network for performance comparison. Although the number of patient is 71 and this is not enough for a deep learning based classification, good results were obtained. Lastly, the effect of radiomic features for the survival time is searched by using classical survival analysis methods such as Proportional Cox Hazar Model, Kaplan Meier Survival Curve and hazard ratios of each covariates. The results shows that 3D radiomic features give better performance than 2D radiomic features. The best results for both classification tasks are obtained by Convolutional Neural Network (CNN).

Keywords: Radiomics, Deep Learning, Survival Analysis, Metastatic Melanoma, Lifex, Transfer Learning

1. Introduction

Metastatic Melanoma (MM) is the type of malignancy that has the highest mortality rate (Sabaila et al., 2015). Melanoma causes more than 10,000 deaths each year. It has a 98% five-year survival rate when diagnosed and treated early, at a local stage. In contrary, with advanced melanoma, at a metastatic stage (i.e. with distant secondary lesions), five-year survival rate decreases up to 5 or 19% (Lens and Dawes, 2004). The first breakthrough in the treatment of metastatic melanoma was the use of targeted therapy blocking BRAF and MEK, which can unfortunately be used in only 40% of patients whose tumour present a BRAF V600 mutation (Larkin et al., 2014). The second and major was the introduction of immunotherapy (ipilimumab then pembrolizumab and nivolumab) that proved to be associated with long overall survival in metastatic melanoma (Robert et al., 2015). Immunotherapy is now commonly used as a first lign therapy in MM.
The improvement in survival remains however heterogeneous when treated with immunotherapy. Adverse events can occur during immunotherapy and can be severe (Martin-Liberal et al., 2015). Clinicians need predictors to the response of immunotherapy for each patient to evaluate if he can benefit from immunotherapy of if another treatment should be preferred (targeted therapy, enrollment in clinical trials). The only predictor known currently are the number of metastatic sites (and notably presence of liver, lung and brain lesions)

and the serum level of LDH.
CT scan is one the mostly used mean to assess the extension of the metastatic disease (it can be combined with a brain MRI and/or with a PET CT). Visual analysis by a radiologist of the CT scan can provide with the number of sites with lesion (notably the presence of brain, liver and lung lesion), the size and number of lesions. Figure 1 shows that an CT scan slice of brain metastasis of melanoma.

Figure 1:Brain metastases of melanoma

Diagnostic imaging can provide the information such as tumors overall shape, size, density. Heterogeneity can not be quantified by the human eye, yet it has proven to be a prognostic factor in some tumors (Rao et al., 2016). Texture analysis is an emerging technique that can be applied to quantify tumor heterogeneity (Lubner et al., 2017). This technique analyzes the distribution and relationship of pixel gray levels in the tumor and identifies spatial variation of individual gray levels or patterns.

The use of extracted high-dimensional data from medical images (CT, MRI, PET CT, Ultrasound) is often called in medical literature radiomics. It is mostly used to evaluate prognosis in tumors, to detect or characterize tumors.It is a quickly developing area in medical image analysis/ Figure 2 shows that a standard radiomics based study flow. It first requires annotation (and often segmentation) of the lesion in the medical image. Then radiomics features are extraction for each lesion; they describe the distribution of gray level in the segmented image. Radiomics (imaging) features can then be combined with other data for the patient such as clinical data (age,sex...) and even genomics. Finally those combined radiomic features can be used for classification, regression and survival analysis.

Besides from radiomics, another approach to Computer Aided Diagnosis (CAD) is Deep Learning (DL). It is becoming powerful tool for CAD systems day by day. With increased number of images, classification and regression models in medical image analysis are extremely popular. Deep learning can be seen as a type of machine learning. It is based on artificial neural networks. Learning phase of a deep neural network is either supervised or unsupervised. If a label or groundtruth is in the training step of the network, it is called supervised. Unlike supervised learning, unsupervised learning has no criteria to classify the data, the computer determines the classes. Convolutional neural networks (CNN) are a special type of Deep neural network (DNN) that are used for images.

In the last recent years, it can be said that hand crafted features (Radiomics) and deep features are equally . Figure 3 shows Google scholar articles for deep learning and radiomics features. Although deep learning is very popular in medical imaging field, the interest of radiomics has been increasing. Figure 4 shows that the main 4 applications domain of Hand-crafted and Deep learning-based Radiomics.

In this study, we aimed at comparing radiomics and deep features to predict survival time and therapy response of the metastastic melanoma patients.

Radiomic data is combined with different feature selection methods and machine learning classifiers. Then by using CT scans, 2D tumor patches are extracted to be used in a DNN for the same classification tasks. The main aim is to show the difference between radiomic features with machine learning versus DNN.

This thesis is organized as follows:
Section 2 explains applications and state of art articles of comparison between radiomics and deep learning for survival time analysis, therapy response or tumor classification based on survival time.
Section 3 presents the feature selection, classification, data augmentation algorithms for Radiomics features and convolutional neural network approach for classification in terms of time and therapy response. In addition, survival analysis is explained in the section 3. After overview of the methods, the implementation part is explained in this section as conclusion. Results for each implemented method are in the section 4. Discussion for the results are in section 5. The thesis is concluded in section 6.

2. State of the art

The review takes into account not only the classification by using Radiomics features, but also the deep neural network approaches to predict survival time, therapy response survival analysis. The state of art studies cov-

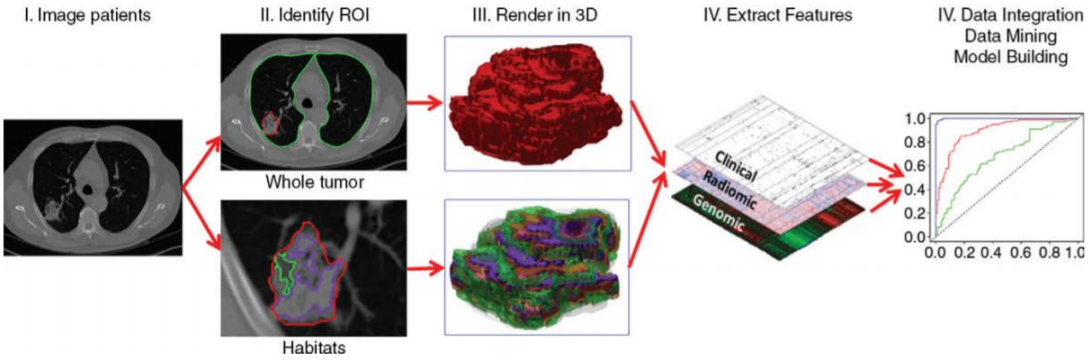


Figure 2:Pipeline and usage of radiomics (Gillies et al., 2016)

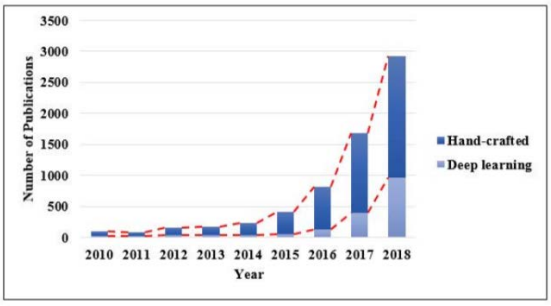


Figure 3: Google Scholar Results (Afshar et al., 2018)

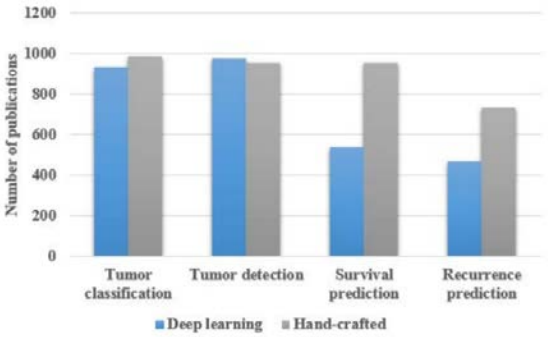


Figure 4: Number of publications in the 4 application domains of radiomics and deep features in medical image analysis (Afshar et al., 2018)

ering the master thesis topic partly are; survival prediction or classification with the comparison of radiomic features and deep features. The literature review mainly focus on the studies that use CT image modality. Existing groups can be divided into 3 and can be explained by selected articles:

2.1. Radiomics Features and Machine Learning

In this section, the radiomic features are combined with machine learning methods. Radiomics features are extracted by using softwares and different feature selection methods are applied to reduce the dimensionality of feature maps. Then conventional machine learning classifiers such as Support Vector Machine, K-nearest neighbor classifier, Random Forest Classifier are utilized for classification in order to predict survival time of the patients.

Sun et al. (2018) extracted 339 radiomics features that are extracted from CT images of 283 patients presenting with a pulmonary cancer (NSCLC). Then 5 feature selection methods and 8 machine learning methods were combined for the Overall Survival (OS) prediction. The metrics for evaluation was concordance index. Method evaluation is based on 3 cross-validation (CV).

Chaddad et al. (2017) proposed a method for NSCLC patients: CT scans of 315 NSCLC patients were used to extract 24 image features. Random forest classifier were used to predict patient survival groups. Log-rank test, Kaplan-Meier estimation and used for feature importance.

Zhang et al (Zhang et al., 2017) studied 112 subjects with lung cancer and the aim is to predict of lung cancer recurrence and death. This study uses only radiomics features.

Aerts et al. (2014) study is one of the most broad study for relationship between radiomics features and survival prediction in terms of number of patients with 1019 patients. In Yin et al. (2019) et al study aimed to find optimal machine learning methods for preoperative differentiation of sacral chorddoma (SC) and sacral giant cell tumour (SGCT) by using CT images. Total number of patient was 95 and 3 feature selection methods (Relief, LASSO and Random Forest) were combined with 3 classification methods (SVM, RF and GLM).

In the above approaches, the number of patients highly correlated with the performance of the classifiers.

2.2. Deep Features and Machine Learning

Some of the following studies use not only deep features solely but also a combination of radiomics and deep features. Aydin et al. (2017) proposed a search for the comparison between radiomic features versus cnn features for overall survival of brain tumour patients. They concluded that the best prediction accuracy can be achieved by using CNN features and linear discriminant

combination. On the other hand Nie et al. (2019) used 3D patches to feed the convolutional neural network to predict the survival time for brain tumor patients. Final classification was achieved by using Support Vector Machines. 3D patches classification is a novel approach but the main limitation is number of images. Lao et al. (2017) proved that the combination of deep learning features and radiomic features performed better than only radiomic features for prediction of survival in an aggressive primary tymor of the brain, glioblastoma multiforme.

2.3. Deep Learning

This approach aims to use only convolutional neural network for survival time prediction classification.
van der Burgh et al. (2017) combined the deep learning features and clinical data to feed the network to predict survival time in terms of classification.
Li et al. (2017) study used the fully CNN pipeline for survival analysis of rectal cancer.
Haarburger et al. (2018) used 2D patches for feature extraction by using Resnet architecture. The deep features were then combined with radiomic features for survival predictions.

Table 1 shows the summary of state-of art review and the information about image modality, number of patients and the feature properties.

It can be noted that studies from Table 1 focus on mainly one type of cancer and on the primary tumor, which means that images come from one part of body. Indeed, the number of patients are much more than this project. Number of samples have an effect on the machine learning and mostly the deep learning studies. Besides, those studies tried to answer only one question at the time : classification in terms of survival time, therapy response or survival analysis. This thesis aimed to find the answer all three questions.

2.4. CAD in metastatic melanoma

Only two studies used CAD in metastatic melanoma, with only radiomic features. Smith et al. (2015) used radiomic features from CT before treatment and their changes on the CT after the beginning of treatment to find prognostic factors of survival. This study contains few patients (23), concerns an out of date treatment (bevazicumab).
Recently, Durot et al. (2019) analyzed radiomic features on CT before pembrolizumab. But this study has limits that this thesis wants to overstep : small number of patients (31), the features were evaluated on multiple lesion by patient but only their mean value was used, the features were extracted on one slice only, and finally no classification nor prediction was performed or evaluated (the result is just an association of features with survival).

3. Material and methods

3.1. Materials

This is a monocentric retrospecticve study. All patients treated with anti PD1 immunotherapy (nivolumab or pembrolizumab) for metastatic melanoma in the Universite Clermont Auvergne Hospital were included, with a total of 71 patients. Images of the last CT before immunotherapy were visually assessed by a radiologist to identify the metastasis. Then each mesurable lesion was manually segmented on the DICOM image by two radiologists with 4 and 8 years experience. It resulted for each lesion in one volumic (3D) ROI for the whole tumour and one 2D ROI for the largest slice of the lesion. The number of lesions per patient varied from 1 to 11; the total number of lesions was 539. Most lesion were pulmonary lesions, then brain lesions.

3.2. Methods Overview

3.2.1. Radiomics Features

Radiomics is a recent field that aims to extract quantitative features from medical images for decision support. It can be defined as the conversion from images to high dimensional data. Combination of radiomics data and clinical data increase the power of decision support systems. Radiomics features can be extracted from tomographic images. These features can be divided into 3 categories: First order, second order and high order radiomics.
First order radiomics take into account the pixel intensity distribution. First order radiomics have 2 categories; shape and intensity features. Shape based features is useful to describe geometrical properties of region of interest. Shape based features are extremely useful for tumor malignancy and therapy response prediction (Afshar et al., 2018). Intensity based features are used to investigate properties of the histogram of tumor intensities. Examples of first order radiomics are compactness, mean intensity, intensity standard deviation, entropy, kurtosis, skewness and uniformity.
Second order radiomics concern texture features. Shape and intensity features can fail when the correlation between different pixels of an image occurs. In that case texture features are most important ones when heterogeneity has an indicator for a tumor type(Afshar et al., 2018). In this thesis, 3 subcategory of texture features are focused on. These are gray level co-occurrence, gray level run length and gray level zone length (GLZLM). Gray level co-occurrence is a matrix to show the frequency of two intensity levels of two neighbour pixels (Afshar et al., 2018). Gray level run length is also a matrix for consecutive pixels. Gray-Level zone length matrix focuses on the size of homogeneous regions in a volume. More details are shown in Table 2. Table 2 shows the radiomic feature categories that are used in this thesis. Table 3 is brief explanation of used radiomic features in this project.

Table 1: Review of Radiomics and Deep Learning articles for CT imaging

Author	Modality	Application Domain	Features	Number of patients
Napel et al. (2018)	CT	Lung Cancer Survival Prediction	Radiomics Features	288
Haarburger et al. (2018)	CT	Image Based Survival Prediction of Lung Cancer Patients	Deep Features	422
Li et al. (2017)	CT+PET	mage basedsurvival analysis of rectal cancer	Deep Features	84
Nie et al. (2019)	Mri+fMri+DTI	Survival Time Prediction of Brain Tumor Patients	Deep Features	69
van der Burgh et al. (2017)	MRI	Prediction of survival on MRI in amyotropic lateral sclerosis	Clinical data+Deep Features	135
Truhn et al. (2018)	MRI	Classification of lesions at MRI	Radiomics vs Deep Features	447
Bibault et al. (2018)	CT	Therapy response for advanced rectal cancer	Radiomics vs Deep Features	95
Chato et al. (2017)	MRI	Prediction of overall survival of brain tumor patients using MRI images	Radiomics vs Deep Features	163
Lao et al. (2017)	Multi modality MR images	Prediction survival in glioblastoma multiforme	Radiomics and Deep	75
Zhang et al. (2017)	CT	Prediction of lung cancer recurrence and death	Radiomics	112
Aerts et al. (2014)	CT	Lung ,head and neck cancer survival prediction	Radiomics	1019
Oikonomou et al. (2018)	CT and PET	Lung cancer survival prediction	Radiomics	150
Oakden-Rayner et al. (2017)	CT	Longevity prediction	Radiomics,deep features	48
Paul et al. (2016)	CT	Lung Cancer short/long term survival prediction	Radiomics and Deep features	81
Wu et al. (2019)	CT	Bladder Cancer treatment response prediction	Deep Features	123

The radiomic feature extraction can be done in different ways: there are softwares such as LIFEx software, 3DSlicer or pyRadiomics python package. LIFEx is a software that is contributed by different universities and institutes in France (Nioche et al., 2018). Compared to pyRadiomics, LIFEx is user friendly. There is no need to code any line to extract radiomic features, hence used by radiologists; manual segmentation can also be performed on LIFEx. The radiomic feature extraction in this work was done using LIFEx. LIFEx enables the calculation of conventional, histogram-based textural and shape features from all image modalities (Nioche et al., 2018). A total of 38 radiomic features were extracted from all ROIs.

3.2.2. Feature Selection and Classification Algorithms

In the decision support and computer aided diagnosis systems, the accuracy depends on the amount of data. In the Gillies et al. (2016) study, rule of thumb is 10 patients per feature for the binary classification case. Since the number of patients in this study is 71 and the number of extracted features are 38, feature selection is a critical step to get better results. Feature selection methods are divided into 3 categories. These are filter methods, wrapper methods and embedded methods. Figure 5 shows the block representation of these methods.

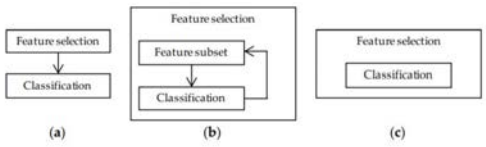


Figure 5: (a)Filter methods, (b)Wrapper methods, (c)Embedded methods

1. Filter methods: These methods are also called univariate methods. They consider the relationship between features and class labels. Redundancy is not considered.
2. Wrapper methods: Wrapper or multivariate methods select the features iteratively by maximizing prediction accuracy of the classifier.
3. Embedded methods: They are very similar to wrapper techniques since they are likewise used to streamline the target capacity or execution of a learning calculation or model. The distinction to wrapper strategies is that a natural model structure metric is utilized amid learning.

In this study total 5 different feature selection algorithms were utilized with radiomic features. These are SFS (Sequential Forward Selection), Boruta, Relief, Recursive and Random Forest feature selection algorithms. Description of the each feature selection method is as follows:

1. SFS: Input of SFS algorithm is whole feature set. Output is a subset of features. Number of selected features k are tuned by the rule of $k > d$ where d is the number of features. It is a wrapper feature selection method.
2. Boruta Feature Selection: It is a wrapper built around the random forest classification algorithm. It tries to find important features in a dataset with respect to the labels. First step is the duplication of the feature vector, and shuffling values in each column. These values are called shadow features. Second step is to train a classifier on the dataset. Usually random forest or logistic regression classifiers are used. Next step, boruta algorithm checks the higher importance features. The condition is that an original feature has a higher Z-score than not only maximum Z-score of its shadow features but also higher than the best of the shadow features. If it is true, then this feature as selected. Selected feature is called a hit. All the hits are found iteratively. Note that a Z-score is the number of standard deviations from the mean a data point. In each iteration, the algorithm compares the Z-scores of shuffled shadow features and the original features to see if the original features are performed better than shuffled shadow features.If it performs better, the algorithm will mark the feature as important. Briefly it can be said that boruta is trying to validate the importance of the feature by comparing with random shuffled copies, which increases the robustness.

Table 2: Radiomics feature types

Category	Description	Sub-category
First Order Radiomics	Information of pixel intensity distribution	
Shape Features	Geometric shape of region	ROI,Sphericity,Compactness, Total Volume, Surface Area, Diameter,flatness and surface to volume ration [2,25]
Intensity Features	Obtained from histogram of the region	Mean Intensity, Intensity Standard Deviation, Median Intensity, Minimum of Intensity, Maximum of Intensity, Mean of positive intensities, Uniformity, Kurtosis, Skewness, Entropy [2,25]
Second Order Radiomics	Information of texture features and relationship between pixels and tumor heteregonity	
Gray Level Co-occurence (GLCM)	Number of times of two intensity levels in a pixel pair. Distance between pixels can be specified .	Contrast,Energy, Correlation, Homogeneity, Variance, Autocorrelation,Dissimilarity, Correlation.
Gray Level Run-Length(GLRLM)	Information from size of homogenous zones	Gray-Level nonuniformity,

Table 3: Radiomics features

Name of the radiomics feature	Category	Brief Explanation
GLCM_Homogeneity	Texture Features	Homogeneity of grey-level voxel pairs.
GLCM_Energy(Uniformity)	Texture Features	Uniformity of gray-level voxel pairs.
GLCM_Contrast(Variance)	Texture Features	Local Variations in GLCM
GLCM_Correlation	Texture Features	Linear dependency of grey levels in GLCM
GLCM_Entropy_log10 and GLCM_Entropy_log2	Texture Features	Randomness of gray-level voxel pairs
GLCM_Dissimilarity	Texture Features	Variation of gray level voxel pairs
NGLDM_Coarseness	Texture Features	Level of spatial rate of change in intensity
NGLDM_Contrast	Texture Features	Intensity difference between neighbour regions
GLZLM_SZE	Texture Features	Long homogeneous zones in an image
GLZLM_LZE	Texture Features	Long homogeneous zones in an image
GLZLM_HGZE	Texture Features	Distribution of the low or high grey-level zone
GLZLM_SZHGE	Texture Features	Distribution of the short homogeneous zones with low or high grey-levels
GLZLM_LZHGE	Texture Features	Distribution of the long homogeneous zones with low or high grey-levels
GLZLM_GLNU	Texture Features	Nonuniformity of the grey-levels or the length of the homogeneous zones
GLZLM_ZLNU	Texture Features	Nonuniformity of the grey-levels or the length of the homogeneous zones
GLZLM_ZP	Texture Features	Homogeneity of homogenous zones
GLRLM_SRE	Texture Features	Distribution of the short or the long homogeneous runs in an image.
GLRLM_LRE	Texture Features	Distribution of the short or the long homogeneous runs in an image.
GLRLM_HGRE	Texture Features	Distribution of the low or high grey-level runs
GLRLM_SRHGE	Texture Features	Distribution of the low or high grey-level runs
GLRLM_LRHGE	Texture Features	Distribution of the long homogeneous runs with low or high grey-levels
GLRLM_GLNU	Texture Features	Nonuniformity of the grey-levels or the length of the homogeneous regions
GLRLM_RLNU	Texture Features	Nonuniformity of the grey-levels or the length of the homogeneous runs
GLRLM_RP	Texture Features	Homogeneity of the homogeneous regions
minValue	Texture Features	Minimum pixel value of the ROI
meanValue	Texture Features	Average of pixel values
stdValue	Texture Features	Standard deviation of pixel values
maxValue	Texture Features	Maximum pixel value
CONVENTIONAL_TLG (mL)	First Order Features	Total Lesion Glycolysis inside the ROI
HISTO_Skewness	First Order Features	Asymmetry of the grey-level distribution in the histogram
HISTO_Kurtosis	First Order Features	Shape of the grey-level Distribution (peaked or flat) relative to a normal distribution
HISTO_Entropy_log10	First Order Features	Randomness of the distribution
HISTO_Entropy_log2	First Order Features	Randomness of the distribution
HISTO_Energy	First Order Features	Uniformity of the distribution
SHAPE_Volume (mL)	First Order Features	Volume of ROI in mL
SHAPE_Volume (# vx)	First Order Features	Volume of ROI in mL
SHAPE_Sphericity	First Order Features	Sphericity of the volume. 1 for a perfect sphere.
SHAPE_Compacity	First Order Features	Compactness of the ROI

3. Relief Feature Selection: It is a filter style feature selection algorithm. It takes a data set with n samples (patients) of p features. The labels of samples are known. The features should be normalized in the interval [0 1]. The number of iteration of algorithm is called m. Algorithm starts with a weight vector (W) of zeros. The length of this vector is also p. Relief takes the feature row vector (X) which belongs to one random sample (patient) and the feature vectors of the sample closest to X (by Euclidean distance) from each class. The closest same-class sample is called 'near-hit', and the closest different-class sample is called 'near-miss'. The weight vector is updated as:

$$W_i = W_i - (x_i - \text{nearHit}_i)^2 + (x_i - \text{nearMiss}_i)^2$$

Thus the weight of any given feature decreases if it differs from that feature in nearby instances of the same class more than nearby instances of the other class, and increases in the reverse case.
At the end of iterations, divide each element of the weight vector by the number of iteration. Result is called relevance vector. Features are selected if their relevance is greater than a threshold T.
4. Recursive Feature Selection: Recursive feature elimination (RFE) is a feature selection method that fits a model and removes the weakest feature (or features) until the specified number of features is reached. Features are ranked by the models coefficient or feature importances, and by recursively eliminating a small number of features per loop, RFE attempts to eliminate dependencies and collinearity that may exist in the model.
5. Random Forest Feature Selection: Random Forest Classifier consists of many decision trees. By using random forest classifier, the features can be ranked in terms of importance. In the random for-

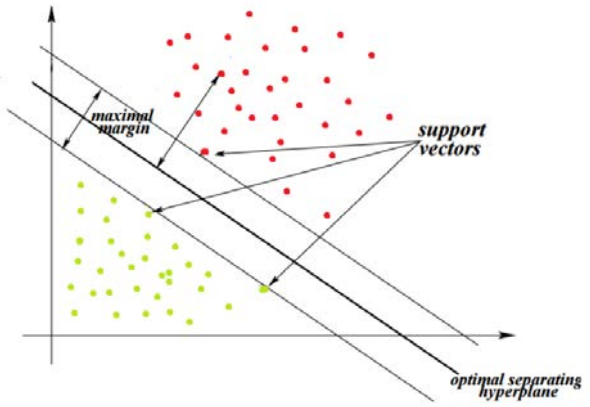


Figure 6: Support vector machine

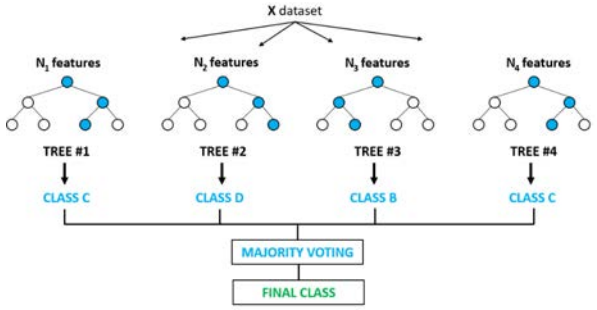


Figure 7: Random Forest Classifier(Adapted from Global Software Support)

est, there is a nice property: for each feature there is a condition at every node in the decision trees. It is designed to split the dataset for classification. Optimal case for splitting the dataset is called impurity. When random forest classifier is trained, it detects the amount of decreased special impurity weight. Since each feature has own special impurity, the features can be ranked according to special impurity.

For the classification, Support Vector Machines, Logistic Regression, Random Forest and k-neighbor classifiers are used.

1. Support Vector Machine (SVM): SVM is a supervised machine learning algoritm. It aims to find a best hyperplane to split a dataset into two classes. A support vector is a data point that is nearest to the hyperplane. They are accepted as critical element of a data set. The term margin is defined as the distance between support vector and hyperplane. SVM choses a best hyperplane with the biggest margin between the hyperplane and any point within the training set. This increases the classifier performance. Figure 6 shows a SVM classification.

2. Random Forest: Random forest consist of many decision tree blocks. A decision tree can be defined as a decision support tool that uses a tree-like graph or model of decisions and their possible outcomes, including chance event outcomes, resource costs, usage. Since a random forest classifier are made by decision trees, each decision tree in the classifier considers a random subset *Ni* of features when forming questions and only has access to a random set of the training data points. This increases the power of random forest classifier. Figure 7 is the representation of random forest classifier. For each feature, there is one decision tree.
3. K-nearest neighbour(KNN): It is also a supervised machine learning algorithm. The algorithm assumes that similar things are close to each other. Therefore, for KNN distance or closeness is the metric for similarity. It calculates the distances between points on a graph.

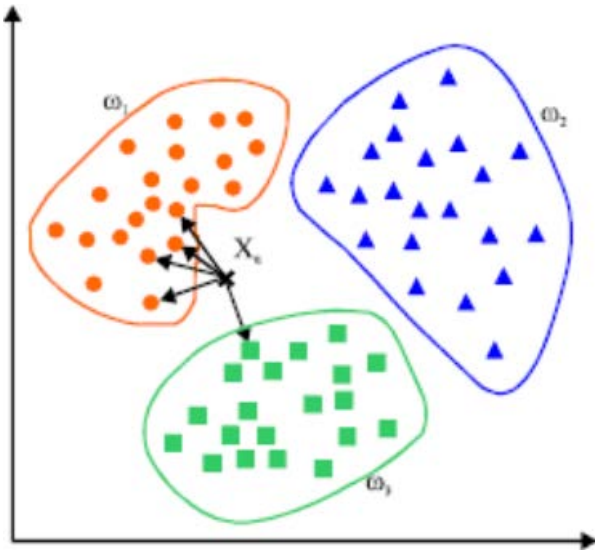


Figure 8: KNN Classifier (Adapted from mathwork)

4. Logistic Regression: It is a supervised machine learning classification algorithm that uses sigmoid function σ. An equation can be written as :

$$h_\theta(x) = \sigma(Z)$$

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

The logistic function has asymptotes at 0 and 1, and it crosses the y-axis at 0.5. Figure 9 shows an example of logistic regression classifier.

All these classification and feature selection algorithms have been applied to radiomic features in the state of art studies. There is no study in the state of art (Section 2) that focused on metastastic melanoma,

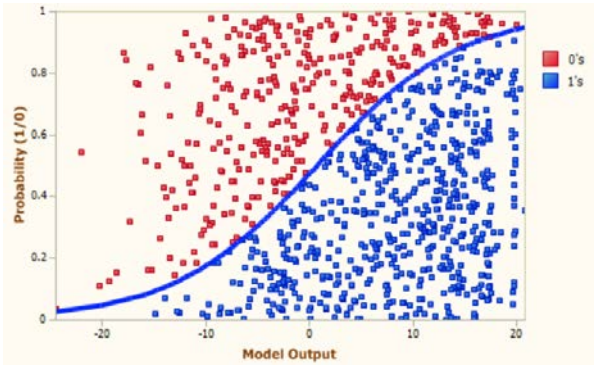


Figure 9: Logistic Regression Classifier(Adapted from *Medium*)

that is the reason why this study aimed to investigate the effect of classification and different feature selection methods on radiomic features from metastatic melanoma lesions.

Table 4 shows that the selected features with different feature selection algorithms.

3.3. Data Augmentation and Data Preprocessing

3.3.1. SMOTE

This study investigates the survival time prediction and therapy response prediction of the patients. Classifier performance depends on data balance. For both cases, there is an imbalanced data problem. Patients are not equally distributed between classes in each study. Imbalanced dataset is the most common accurance for cad and decision support systems due to smaller number of labelled images.

Since the data is real and limited, it is not easy to find a dataset that has 2 classes equally distributed. Therefore, data augmentation to overcome imbalanced datasets are highly important and common in data science. One of the data augmentation method is SMOTE. It is the abbreviation of Synthetic Minority Oversampling Technique. This method is a very popular oversampling method that was proposed to improve random oversampling but its behavior on high-dimensional data has not been thoroughly investigated (Lusa and Others, 2013).

From the Figure 10, SMOTE can be summarized as follow : SMOTE finds the n-nearest neighbors in the minority class for each of the samples in the class.Then it draws a line between the the neighbors and generates random points on the lines.

In this study, SMOTE is used for radiomic features. Many radiomics papers employ SMOTE data augmentation method to overcome imbalanced data solution.

3.3.2. Image Preprocessing and Patch Extraction

Above subsection explains the data augmentation for radiomic features. Recall that in this study total 71 patients' CT scans were used. For classification, deep learning needs a lot of data. When the dataset gets

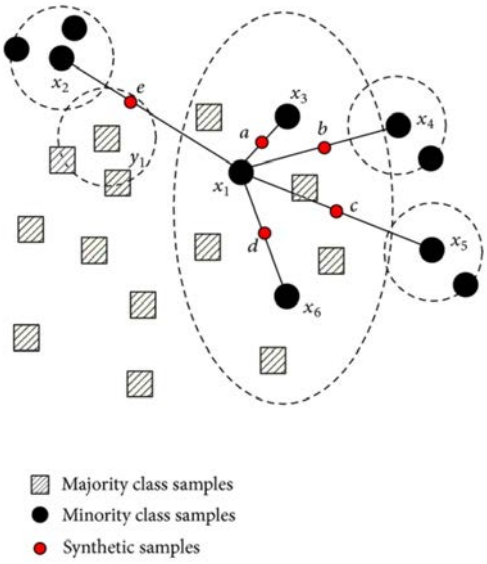


Figure 10: SMOTE Data Augmentation (Lusa and Others, 2013)

larger, the network is able to learn the features and parameters more accurately. Larger dataset makes the deep learning model more generalized for predictions and it prevents the over-fitting. There are 2 main advantages of CT images in this study. First advantage of CT images is that they are 3D volumes. The 2D slices can be extracted easily and augmentation can be done easily. Second advantage is that these 2D slices can be extracted not only from axial axis but also from sagital or coronal axes. In medical imaging field, even though the number of instances are less, data augmentation can be done successfully. Figure 11 is a patient's metastatic melanoma lesion in axial sagital and coronal axes view.

Before 2D patch extraction, all DICOM files were converted to nii files. Nii file type is easier to use in Matlab. Then image normalization is applied using the following equation:

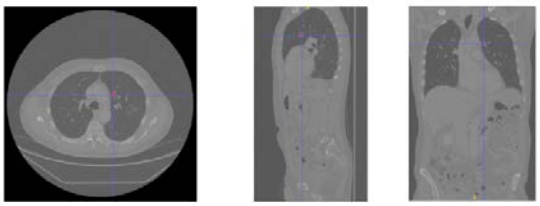


Figure 11: Axial-Coronal-Sagittal Axes of a Tumor Patient

$$I_{new} = \frac{I_{current} - I_{mean}}{Standard\ deviation} \quad (1)$$

where I is the intensity of the pixel.

Data normalization is an important step in this study because CT images are from different parts of body.

Table 4: Selected Radiomics Features for different feature selection methods

Classification Task	Feature Selection Method	Selected Features
Survival Time Prediction	Boruta	'GLRLM_SRHGE', 'GLZLM_ZLNU', 'SHAPE_VOLUME', 'MAXVALUE', 'GLZLM_GLNU'
Survival Time Prediction	Relief	'GLCM_Entropy_log2', 'GLCM_Dissimilarity', 'GLRLM_LRE', 'GLRLM_LRHGE', 'GLZLM_LZHGE'
Survival Time Prediction	SFS	'stdValue', 'GLCM_Correlation', 'GLCM_Dissimilarity', 'GLRLM_SRHGE', 'GLRLM_GLNU', 'NGLDM_Contrast', 'GLZLM_LZHGE', 'GLZLM_ZP'
Survival Time Prediction	Recursive	'HISTO_Skewness5', 'SHAPE_Volume10', 'SHAPE_Sphericity', 'GLRLM_SRE', 'GLCM_Energy', 'GLRLM_SRHGE', 'GLRLM_RLNU', 'NGLDM_Coarseness','NGLDM_Contrast','GLZLM_ZLNU'
Survival Time Prediction	Random Forest	'minValue', 'meanValue', 'stdValue', 'maxValue', 'CONVENTIONAL_TLG (mL)', 'HISTO_Skewness', 'HISTO_Kurtosis', 'HISTO_Energy', 'SHAPE_Volume (mL)', 'SHAPE_Sphericity (only for 3D ROI (nZ>1)', 'GLCM_Correlation', 'GLRLM_HGRE', 'GLRLM_SRHGE', 'GLRLM_GLNU', 'GLZLM_HGZE', 'GLZLM_ZLNU', 'GLZLM_ZP'
Therapy Response	Boruta	'GLZLM_HGZE', 'HISTO_Skewness' ,'meanValue', 'GLRLM_HGRE', 'SHAPE_Sphericity')
Therapy Response	Relief	'GLCM_Correlation', 'GLCM_Entropy_log10', 'GLRLM_SRHGE', 'GLRLM_GLNU', 'NGLDM_Contrast', 'GLZLM_SZHGE'
Therapy Response	SFS	'minValue', 'meanValue', 'CONVENTIONAL_TLG (mL)', 'HISTO_Kurtosis', 'HISTO_Entropy_log10', 'HISTO_Energy', 'SHAPE_Volume (mL)', 'SHAPE_Compacity only for 3D ROI (nZ>1)', 'GLCM_Energy', 'GLCM_Dissimilarity', 'GLRLM_SRE', 'GLRLM_GLNU', 'GLRLM_RLNU', 'GLRLM_RP', 'GLZLM_SZE', 'GLZLM_ZP
Therapy Response	Recursive	'minValue', 'meanValue', 'stdValue', 'maxValue', 'HISTO_Skewness', 'HISTO_Kurtosis', 'HISTO_Energy', 'SHAPE_Volume (mL)', 'SHAPE_Sphericity (only for 3D ROI (nZ>1)', 'GLCM_Correlation', 'GLRLM_LRHGE', 'GLZLM_SZE', 'GLZLM_HGZE', 'GLZLM_SZHGE', 'GLZLM_ZP'
Therapy Response	Random Forest	'minValue', 'meanValue', 'stdValue', 'maxValue', 'HISTO_Skewness', 'HISTO_Kurtosis', 'HISTO_Entropy_log10', 'HISTO_Energy', 'SHAPE_Volume (mL)', 'SHAPE_Sphericity (only for 3D ROI (nZ>1)', 'GLCM_Correlation', 'GLRLM_SRE', 'GLRLM_HGRE', 'GLRLM_SRHGE', 'GLZLM_SZE', 'GLZLM_HGZE'

Equation (1) helps to normalize intensity level of ct images.

In this study, 2D patch extraction is done by using mask window method. For all CT images, ROI masks are available. These are also 3D volumes. Given the dimension of the CT images and the ROI of the lesions, it would not be wise to use the full image as a direct input for the network training. In that regard, each image is split into 30 by 30 pixels windows. The window moves first horizontally and from top to bottom. The starting point (top left or top right) is chosen according to the image orientation. The window is then analyzed to define if there is a full lesion inside this 30 by 30 pixels windows. Bounding boxes of each tumor slices are calculated then the maximum bounding box size is taken for this study. It is obtained as 30 by 30 pixel windows. The bounding box is calculated according to axial axis of each tumor, the reason is that axial axis tumor ROI for all lesions are bigger than coronal or sagital axes.By the augmentation, dataset of 71 patients are enlarged to 21190 . Each CT image is sliced to obtain 2D grayscale images. Since each patient has different numbers of CT scans, the 2D image acquisition is achieved as figure 13. An example of axial-coronal and sagital views of a lesion is shown in Figure 12.

The tumor patch extraction algorithm is detailed in Algorithm 1. This algorithm is implemented in Matlab.

Figure 12: 2D axial-coronal-sagittal patch extraction example

3.3.3. Convolutional Neural Networks

Convolutional Neural Networks (CNN) are one of the best architectures for classification of images or detection a region in the images. CNNs are getting popular everyday due to the proved success. Imagenet challenge is a proof for the power of deep learning. This annual challenge shows that deep learning performance is exceeding the humans (Zaharchuk et al., 2018). The real power of a CNN comes from convolutional layers. They are responsible for the extraction of features across the input image through convolution. Each convolutional layer is connected to a small chunk of spatially connected neurons. Another type of layers in CNN is max pooling. This layer is derived from convolutional layers to reduce the size of feature maps. This helps a lot for reducing the computational time. Fully-connected layers are called flatten layers. Each neuron of the flatten layer is connected to activation function of its previous layer. The main contribution of flatten layers is to learn non-linear relationship in the featuremap. At the end , usually either a soft-max layer or sigmoid layer ends the architecture. Its functionality is to normalize the output.

The main power of CNN comes from discriminant feature extraction. Although CNN is a good approach in medical image analysis, it has some drawbacks. First of all, a huge amount of data is required. This is the first

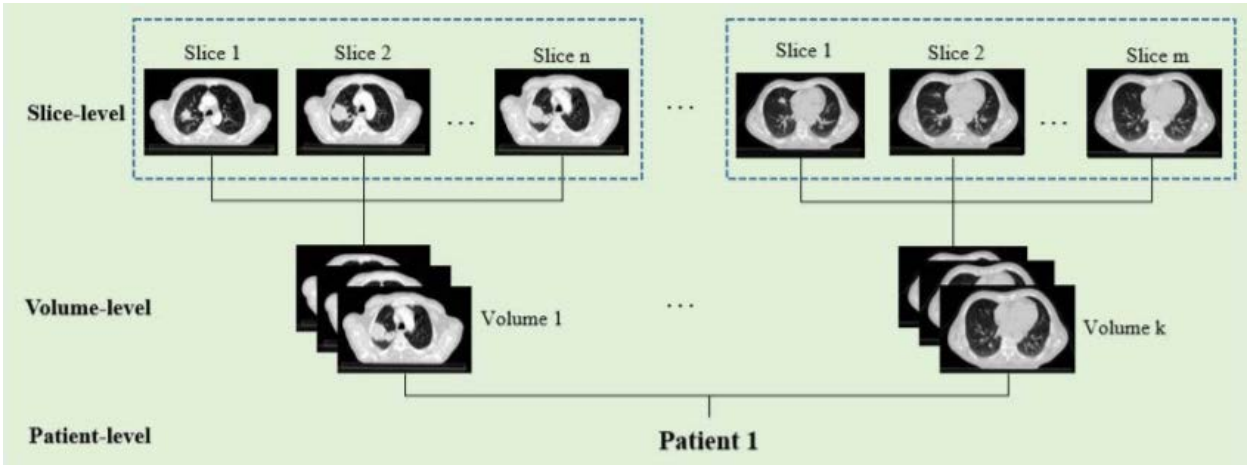


Figure 13: 2D image acquisiton for one patient for axial axis (Afshar et al., 2018)

challenge in medical imaging because annotated data is not much. Data acquisition is expensive and annotation is time consuming. Second drawback is parameter tuning. Deep learning has many parameters and when train a CNN, overfitting is a common issue. In conclusion, it can be said that training of a CNN is very challenging for medical imaging field (Pan and Yang, 2010). The training problem is solved easily. Networks that are trained by Imagenets are available. This method is called $TransferLearning$

It has been proven that transfer learning is more robust than training from scratch (Tajbakhsh et al., 2016).

In this thesis one of the popular CNN architecture is implemented for time prediction classification and therapy response classification of patients by using transfer learning. VGG-16 is network that has 33 convolutional layers a 16 weight layers net. It is built by Oxford's visual geometry group (Simonyan and Zisserman, 2014). Although VGG-16 pretrained network is trained by ImageNet database and these images are not medical images, it is shown that transfer learning can be adapted easily to medical imaging. The network is shown in figure VGG16. The reason behind VGG-16 is that among pre-trained architectures, VGG-16 has the minimum input size. It was trained 48 by 48 RGB images. The architecture is shown in Figure 14.

3.3.4. Deep Features

Traditionally extracted feature methods are called handcrafted feature engineering. Radiomic features are in this category. Although handcrafted features are quite popular in medical imaging, the number of features per image is limited. For example, in radiomics there is no more than 200 features. The main contribution and advantage of deep learning is automatic feature extraction. Until fully connected layers, convolutional layers are responsible to extract big-dimensional features. The power of deep features is that it is not effected by any human bias.

3.3.5. Survival Analysis

Survival analysis is the study of time-to-event data for an individual. In medicine, the most common events are death, therapy response. The main problem for survival analysis in medicine is the time indicator. While for some patients time is the time-to-event, for others is the time of last follow-up. In the second case data are called censored data and usually are shown as 0. Indicator 1 means that the patient died and indicator 0 means that the patient left the study and the last state is unknown.

Survival analysis can show the prediction of hazard. The most common hazard prediction is called "cox proportional hazard model" (CPH) (D.R.Cox, 1972). This model utilizes the predictors (covariates) and it determines each patient's hazards. The output of the CPH is the patient's risk of dying at a time t.

$$h(t) = h_0(t).exp(\beta^{T.cov_i}) \tag{2}$$

In the equation 2, $h_0(t)$ is called the baseline hazard ratio. The risk is symbolized in exponential part of the equation. It assumes that risk is linear combination of the predictors. The relationship between the hazard and the survival time is inversely proportional. If S(i) and S(j) denote the survival time of the patients then i and j relationship becomes as follow:

$$S(i) > S(j) \rightarrow h(j) > h(i) \tag{3}$$

Note that S(i) and S(j) are *concordant* pairs. The most common metrics in survival analysis is $concordanceindex$. It is defined as the ratio of number of concordant pairs and total possible pairs.

$$\text{c index} = \frac{\text{number of concordant pairs}}{\text{total possible pairs}} \tag{4}$$

A Kaplan Meier curve (Leger et al., 2017), shows the probability of survival S(t) in given points of time.

$$S(t) = \frac{\text{Number of patients survived until t}}{\text{Number of patients at the beginning}} \tag{5}$$

Algorithm 1: Patch Extraction Algorithm

```
Input: Input: CT image and mask
height=30, weight=30, stride=30
 [rowsintensity,colsintensity,numberofslice]=dimensions of CT
 for i ← 1 to numberofslices do
    intensity=CT(:,:,numberofslice)
     mask=mask(:,:,numberofslice)
     f=sum(sum(mask));
     if f≠ 0 then
    for i ← 1 to rows do
        Finalrow=initialrow+height;
         if Finalrow<rowsintensity then
        for i ← 1 to colsintensity do
            Finalcolumn=initialcolumn+width;
             if Finalcolumn<colsintensity then
            newmask= crop(mask,[initial initialcolumn height-1 height-1]);
             e=sum(sum(newmask));
             if e = f then
            newintensity=crop(intensity,[initialrow initialcolumn height-1 height-1]);
             newintensity=resize(intensity,[48,48]);
             newmask=resize(mask, [48 48]);
            initialcolumn=initialcolumn+stride;
        initialcolumn=1
    initialrow=initialrow+stride;
    _initialrow=1;
Output: Output: 2D 48x48 patches
```

But in medicine, clustered survival data are widely common in clinical studies. The most important property of a clustered data is that observations from different patients are independent, while observations within a patient may be correlated(LIANG and ZEGER, 1986). To apply cox proportional hazard model on clustered data, a set of estimating equations is provided by the standard partial likelihood equations that incorrectly ignore within cluster. To overcome clustered data Royall (1986) proposed generalized estimated equations(GEE). It is used for estimation the parameters (covariates) of a linear model with unknown correlation between outcomes. In R programming, Cox proportional hazard model can approach GEE by adding *cluster* parameter to a CPH model (Xiaohong Zhang, 2006). In this thesis, CPH used with *cluster* parameter in order to approach the problem as GEE for solving the correlation problem of clustered data.

3.3.6. Therapy response and the survival time classification

In this study, therapy response levels are obtained from clinical data. For the survival time classification, the clinical data has also the time between 2-follow up for each patient. This information is used and assumed that patient lived between these 2-follow up. The classification is binary for both cases. Therapy response labels are bad or good prognosis and the survival time classification labels are patients who lived in more than 1 year and less than 1 year after treatment initiation.

3.4. Implemented Methods

3.4.1. Machine Learning and Radiomics Features

In this case, a total 5 feature selection and 4 classification methods were combined for both therapy response and patient survival time prediction classification. And, the effect of data augmentation is also considered. SMOTE data augmentation method is used on 2D and 3D radiomic features. In total $4 * 5 * 2 = 40$ different combinations are applied to the following label and data:

1. Therapy response prediction with 2D radiomic features (40 combinations)
2. Therapy response prediction with 3D radiomic features (40 combinations)
3. Survival time prediction with 2D radiomic features (40 combinations)
4. Survival time prediction with 3D radiomic features (40 combinations)

A total of 160 different combinations were implemented. For each model, 5 k-fold cross validation was used in order to tune the hyper parameters. It is called grid-search and the aim was to determine the optimal values for a given model.

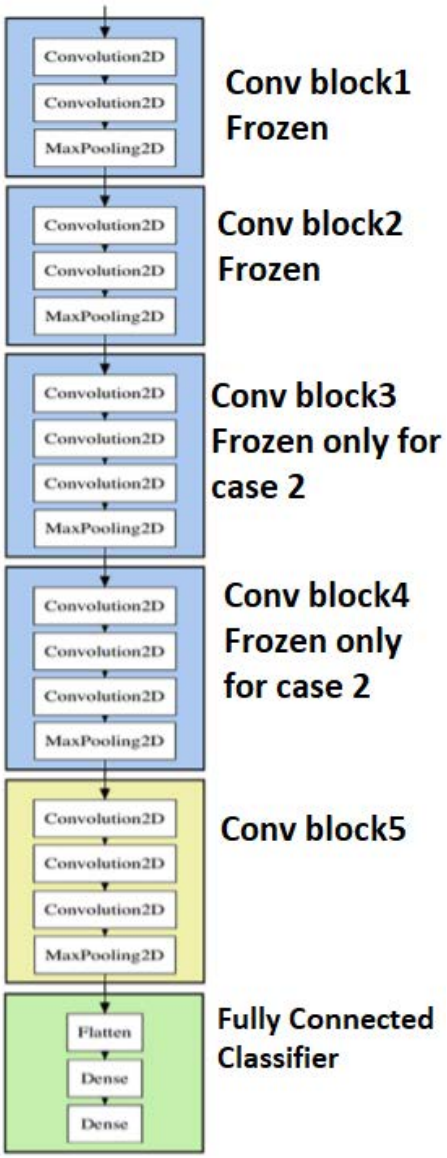


Figure 14: VGG-16 Architecture

In this study, the data type was multiple lesion per patient and the aim was to predict the therapy response and survival time patient based. Therefore, leave one out cross validation was used to assign the label for predictions. For each iteration of a model, one patient was kept as test set and rest is used for training. After training, predictions for each tumor in a patients were assigned. To finalize the patient based assignment, the majority class of the predictions were used. For each patient, the majority of the predictions were assigned as the patient based prediction of the model.

3.4.2. Machine Learning and Deep Features

In this method, extracted patches were used to feed the network. The VGG-16 was used to extract deep features from each patch. Then all the features were stored as .csv file. Every patient has his own *.csv* file. Next step was the feature reduction. Due to huge number of covariates (512), this step was needed. Decision tree feature selection method was used for feature reduction. Number of covariates decreased from 512 to 178. Then SVM, Random Forest, Logistic Regression and K-nearest neighbour algorithms were used to classify the deep features in terms of therapy response and the survival time prediction.

3.4.3. Deep Learning

VGG-16 architecture was used for both classification tasks. Transfer learning was used by freezing the first 4 convolution blocks: Conv block 1 , conv block 2,conv block 3 and conv block 4. Later on, due to huge amount of augmented data, only freezing first two convolutional blocks (conv block 1 and conv block 2) was used for both cases. Since the tasks are binary classification, binary crossentropy is set as loss function and sigmoid activation function is used as the last layer activation function. The method for patient classification is applied in a same way machine learning section. For each patient, subfolders were created. Each subfolder, has the all tumor patches from axial coronal and sagittal axes. In each iteration of the network, one patient was kept as test set and the rest was used as training the network. Number of epochs per patient was 5. At the end of each iteration, network predicted the labels for all the patches in the test folder. Later on, patches labels were voted and majority class label was assigned as the predicted patient label.

In deep learning, activation function is one of the most important element. Currently, the most common activation function is the Rectified Linear Unit (ReLU). First, the ReLU activation functions were utilized in the fully connected layers. After the network fails, other activation functions were searched. In 2017, Google brain proposed a new activation function which is called *swish* (Ramachandran et al., 2017). Swish can be simplified as $f(x) = x * sigmoid(x)$. (Ramachandran et al., 2017). Their experiments show that Swish is a better option than ReLU as activation funcion in DNN across a number of challenging data sets. Swish is as simple and similar as Relu and it is easy to replace Relu with swish in a deep neural network. Swish and relu activation functions are shown figure 15:

The main challenge of this study is survival time classification. The reason is that, all patients have same disease and some of them are censored which is their events are unknown. VGG-16 has not used to classify the patients in terms of survival time. The closest study Haarburger et al. (2018) used the pretrained Resnet-50 as feature extractor then these features are combined with radiomics features.

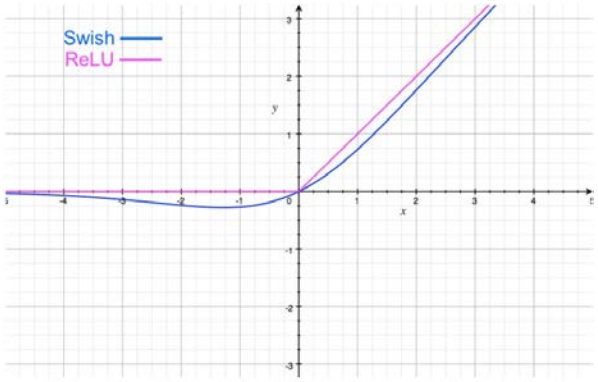


Figure 15: Swish and Relu activation functions (Ramachandran et al., 2017)

For a neural network, an activation function can be written as:

$output = activationfunction(dotproduct(weights, inputs + bias)$

Equation shows that the inputs are multiplied by weights, and a bias value is added to the result. The bias value allows the activation function to be shifted to the left or right, to better fit the data. Hence changes to the weights alter the steepness of the sigmoid curve, whilst the bias offsets it, shifting the entire curve so it fits better. Note also how the bias only influences the output values, it doesnt interact with the actual input data (Lacki, 2017).

The bias is added to the last layer of sigmoid's function and the network is trained with biased last layer. Biased values are different in two tasks. While 0.7 is the new threshold for the patient time classification problem, 0.4 is the new threshold for patient response classification. The effect of bias can be seen in Figure 16. Bias moves the sigmoid function either $-x$ or $+x$ direction. It can be said that the sigmoid moves to $+x$ direction for survival time prediction problem, whereas this function moves to $-x$ direction for therapy response classification. Bias in the output layer is highly recommended if the activation function is sigmoid (Huang et al., 2006).

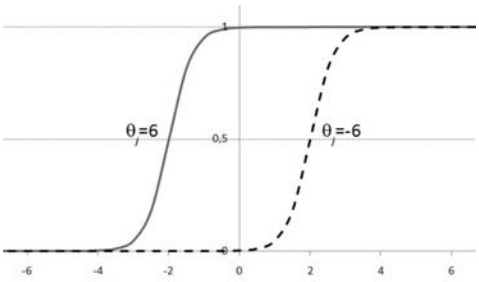


Figure 16: Effect of bias

In summary, fully connected layers of Figure 14 have modified. By adding bias, last layer has a sigmoid activation function with threshold of 0.7 and 0.4 instead of 0.5. Then first two layer's activation functions are set to *swish*.

3.4.4. Survival Analysis

For survival analysis, Kaplan Meier curves for good and bad prognosis patients are plotted. Besides, the hazard ratios for each radiomic covariate is calculated. To investigate the hazard ratios of each radiomics subgroups, first and second order radiomic features are clustered and cox proportional hazard model is calculated.

3.4.5. Evaluation Metrics

The performance of two classifications are evaluated by accuracy and the confusion matrix that are obtained by CNN and machine learning classifiers' predictions.

1. Accuracy Accuracy is the most common and one of the most robust evaluation metric for a classifier performance. Figure 17 shows a confusion matrix. From that matrix, accuracy is calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

2. Confusion Matrix: Performances are calculated on confusion matrix. A confusion matrix keeps the all information about correct and wrong predictions. True Positive, True Negative, False Positive and False Negative values are the elements of that matrix.
Indeed, confusion matrix is not helpful only accuracy; sensitivity and specifity can be calculated by this matrix. In medical imaging field, sensitivity and specifity are as important as sensitivity. Sensitivity is the indicator that how often a test correctly generates a positive result for people who have the illness. In contrast, specificity is a metric for a tests ability to correctly generate a negative result for people who are not sick. Sensitivity and specifiy are calculated as in the equation (7) and (8).

		Prediction outcome: positive	Prediction outcome: negative	
Actual value	positive	TP	FN	$TP + FN$
	negative	FP	TN	$FP + TN$
		$TP + FP$	$FN + TN$	

Figure 17: Confusion Matrix

$$Sensitivity = \frac{TP}{TP + FN} \quad (7)$$

$$Specifity = \frac{TN}{TN + FP} \quad (8)$$

4. Results

In section 4.1, the best 10 performances are shown. The performances of all combinations is in Appendix A.

4.1. Machine Learning and Radiomics Features

Patient Survival Time Prediction : The best 10 combinations in terms of accuracy are shown in Table 5. The best accuracy is achieved by the combination of Recursive feature selection, logistic regression classifier and smote data augmentation with 3D radiomic features.

Table 5: Survival time prediction

Feature Type	Combination	Accuracy	Sensitivity	Specifity
3D	Recursive Feature Selection+Logistic Regression Classifier+Smote Data Augmentation	0.8591	0.9111	0.7692
3D	Relief Feature Selection+ KNN classifier+ Smote Data Augmentation	0.8169	0.7555	0.9230
3D	Random Forest Feature Selection+ Logistic Regression Classifier	0.8169	0.8444	0.7692
3D	SFS Feature Selection+ Logistic Regression Clasifier	0.8169	0.9777	0.5384
3D	Recursive Feature Selection+Logistic Regression	0.8169	0.8881	0.6923
3D	SFS Feature Selection+Logistic Regression + SMOTE	0.8169	0.9777	0.5384
3D	Random Forest+Logistic Regression Classifier+SMOTE	0.8028	0.8881	0.8076
3D	Recursive+KNN	0.7883	0.7111	0.9230
3D	Recursive+KNN+SMOTE	0.7887	0.6666	1.0
3D	SFS+SVM	0.7746	1.0	0.3846

Therapy Response: The best 10 combinations in terms of accuracy are shown in the Table 6. The best result belongs to the combination of SFS selected features, Random Forest Classifier and Smote data augmentation.

Table 6: Therapy response classification with radiomics features

Feature Type	Combination	Accuracy	Sensitivity	Specifity
3D	sfs+rf+yessmote	0.8611	0.9117	0.8157
3D	Boruta feature selection+SVM classifier+ SMOTE	0.8611	0.9411	0.7894
3D	Boruta feature selection+Random Forest Classifier+SMOTE	0.8472	0.9111	0.7894
3D	Boruta feature selection +KNN classifier+ SMOTE	0.8472	0.9111	0.7894
3D	Random Forest Feature Selection+KNN classifier	0.8472	0.7647	0.9210
3D	SFS feature selection+KNN classifier+SMOTE	0.8472	0.8823	0.8157
3D	Borura selection+Random Forest Classifier	0.8333	0.8235	0.8421
3D	Random Forest Feature Selection+KNN classifier+SMOTE	0.8333	0.7647	0.8947
3D	Random Forest Feature Selection+SVM+SMOTE	0.8194	0.7647	0.8684
3D	Relief Feature Selection+Random Forest Classification	0.8194	0.7058	0.9210

4.2. Machine Learning and Deep Features

As it is stated in materials and method section, extracted deep features from 2D tumor patches are used in VGG-16 to extract deep features. Then the conventional machine learning classifiers are classified with leave one out cross validation to classify patients according to 1 year survival time and therapy response. Table 7 is a summary of time prediction classification results and Table 8 shows the results for therapy response classification.

Table 7: Time Prediction-Deep Features

Feature Type	Combination	Accuracy	Sensitivity	Specifity
Deep	Logistic Regression	0.58	0.5869	0.56
Deep	SVM	0.6857	0.60	0.8127
Deep	Random Forest	0.74	0.86	0.5
Deep	KNN	0.6771	0.5813	0.8076

Table 8: Therapy Response-Deep Features

Feature Type	Combination	Accuracy	Sensitivity	Specifity
Deep	Logistic Regression	0.61	0.5579	0.76
Deep	SVM	0.70	0.61	0.83
Deep	Random Forest	0.72	0.86	0.5
Deep	KNN	0.78	0.77	0.63

4.3. Deep Learning

The tables show the performance of VGG16 network on survival time prediction and therapy response classifications on metastastic melanoma patients. As it is explained in the materials and method section, two different approaches are applied to these classification tasks. First one is to train first 4 convolutional blocks of VGG-16 and second method is to train only two convolutional blocks. The result of 4 frozen convolutional blocks VGG-16 is shown in Table 9 and Table 10 shows the 2 frozen convolutional blocks.

Table 9: Result for VGG-16 with 4 frozen blocks

Classification Task	Accuracy	Sensitivity	Specifity
Survival Time Prediction	0.82	0.635	0.747
Therapy Response	0.78	0.80	0.53

Table 10: Result for VGG-16 with 2 frozen blocks

Classification Task	Accuracy	Sensitivity	Specifity
Time prediction	0.92	0.9259	0.8723
Therapy Response	0.901	0.85	0.9333

4.4. Survival Analysis

Kaplan Meier survival curves for each therapy response group are plotted. Cox proportional hazard models were calculated as a whole covariates, only histogram based covariates and only texture features. The c-index of each case was calculated. The cumulative event and cumulative hazard curves were plotted. Hazard ratios were calculated. The population's survival curves for all covariates and only histogram based covariates were plotted. For each radiomics features, the z and p and values were obtained.

Hazard ratio measures of an effect of an intervention on an outcome of interest over time. p value shows the effect of a covariate on survival time. Lower p values indicate that this covariate has a significance on survival time. The coefficient of a covariate is interpreted as the hazard ratio. It indicates the effect of covariates for hazard. Figure 21 shows the all 3D radiomic features' hazard ratios. The last column of the figure corresponds to p values for each covariate. The p values who have $asterisks*$ show that these covariates effect the survival time. Third column and fourth column are for confidence interval. Confidence interval is the precision of the Hazard Ratio. In the figure, fourth column is the visual confidence interval. In this column it is seen that there is a vertical reference line. This line divides the

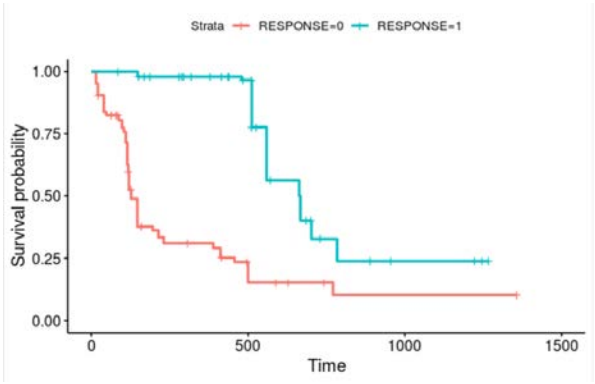


Figure 18: Survival probabilities of each therapy response labeled patients over time

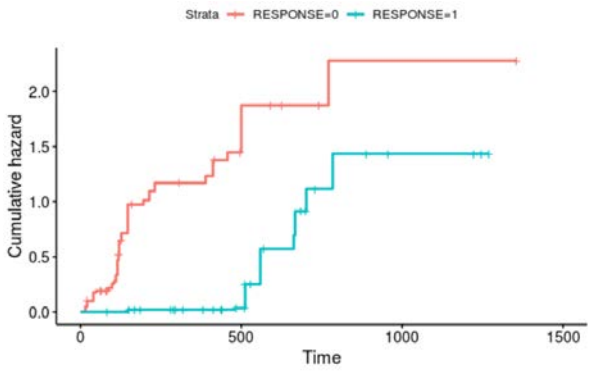


Figure 19: Cumulative hazard curves of each groups

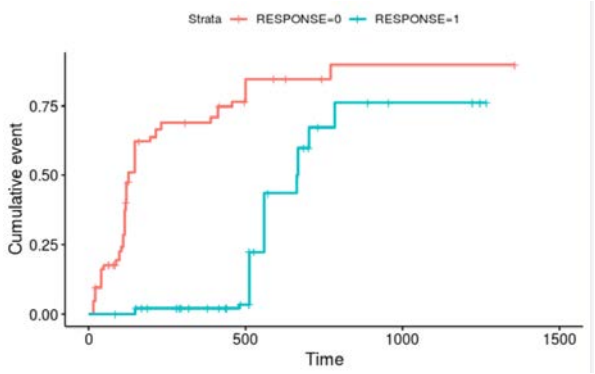


Figure 20: Cumulative event curves of each groups

interval into two. If this line locates on the confidence interval of a covariate, then this covariate is statistically not significant. Even though the p value shows that a covariate has effect on survival time, confidence interval rejects this statement. Thats why confidence interval is the confirmation of p value.

For Cox-proportional hazard model, this study has 3 cases. The first model has all the radiomics features. It is represented in Figure 21. Second model considers only histogram features which are in Figure 22. Last case takes into account only texture features and it is in the Figure 23. The results of three cases is explained as follow: In case one, significant features are shape sphericity, GLZLM_ LZE and GLZLM_ ZLNU but the confidence interval (CI) indicates that even though these covariates' have significance in survival time, it is failed for CI. Figure 22 has one covariate who has a significant p value. For HISTO_Skewness , both p value and CI indicate of significance in a positive way. For the third case, GLRLM_SRE, NGLDM_contrast, GLZM_LZE and GLZLM_ ZLNM have the significant p value. Among these features, only NGLDM_contrast pass the CI test.

Survival curve for each therapy response group is plotted. Indeed, the cumulative hazard and cumulative event graphs are shown in below.

Integral of the hazard function is called cumulative hazard function.In other words, it gives the probability of failure at time x given survival until time x. Figure 19 is the cumulative hazard functions of this study. When proportion of patients that die in one year versus time is plotted, figure 20 is obtained. It shows that the proportion of patients that die at time t.

5. Discussion

5.1. Radiomics and Machine Learning

In radiomics, the results are surprisingly good. First glimpse is that 3D radiomic features are much better than 2D radiomic features. This is an expected result according to Ng et al. (2013) study. For the feature selection algorithms, even though the selected features by SFS are the most meaningful covariates with the combination of random forest classifier, it is obvious that Boruta and Random Forest feature selection algorithms are the most powerful 2 feature selectors for both classification tasks. With the classification algorithms, logistic regression works in high performance for survival time classification. On the other hand, KNN works surprisingly good for therapy response classification task. The effect of data augmentation and balanced datasets are shown in the survival prediction table as inverse proportional. Another example is from therapy response table. In the table 6, the random forest features KNN classifier combination has a better performance than random forest features, KNN classifier and SMOTE. It can be said that synthetic data augmentation does not always increase the performance of a classification task. Usually, data augmentation helps to increase the accuracy, SMOTE is not efficient in a positive way with the random forest covariates. But, it is seen that for both tables, the highest accuracies are achieved by positive effect of SMOTE.

5.2. Deep Features and Machine Learning Classifier

With the high dimensional deep feature data, it can be said that a traditional classifier failed. But, the reason behind the fail of deep features can be explained. Deep

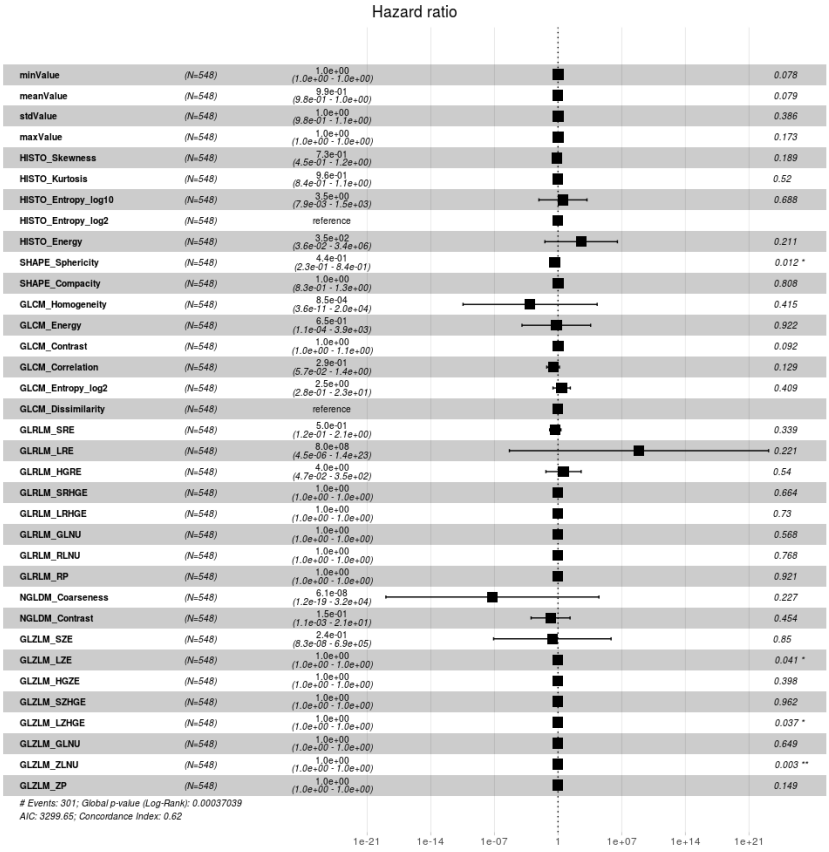


Figure 21: Forest plot for a Cox-regression model fit.

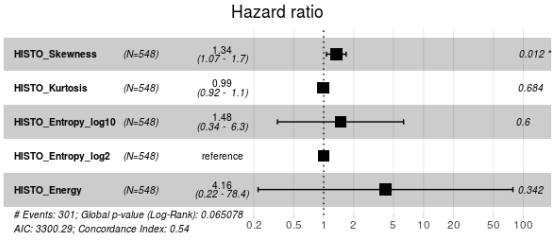


Figure 22: Forest plot for the cox model who has only histogram based covariates

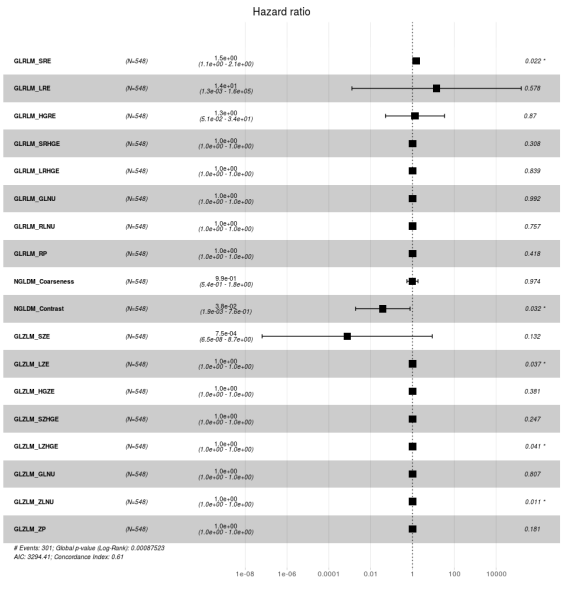


Figure 23: Forest plot for the cox model who has only histogram based covariates

features are extracted from 2D tumor patches, it can be said that all the deep features are 2D features. When 2D radiomic features and deep features performance are compared, it can be seen that results are not extremely different. It can be concluded as 2D features and traditional classification algorithms do not work properly. Although annotation of 3D tumor volume takes much more time than one slice tumor annotation, according to the results, annotations through whole tumor is worth it.

5.3. VGG-16

This study shows that changing the activation function and adding some bias to sigmoid function can be a novel approach for survival time prediction. Until now survival analysis as classification problem has not been used by VGG16. Without clinical data, pure deep features and deep learning classifier are the best combina-

tion in this study. Note that the number of patient in this study is 71. Although the total number of lesions is 539, it is known that the data is not sufficient to feed a deep neural network. Since the tumors from same patient may have the same information, there is few variety of data. All these reasons are solved with changing the sigmoid thresholding. The threshold is not a specific number, it increased only from 0.5 to 0.7 for time classification task and it decreased from 0.5 to 0.4 for therapy response. If a sufficient number of patient was supplied to the project, the deep learning results would be much better with no biased activation function. Indeed, it is known that variance of an activation function can cause overfitting whereas bias can cause underfitting a network (**?**). It can be said that for these classification tasks, overfitting probability is very low.

The effect of swish activation function is seen in this study. When the VGG-16 last layer had Relu activation funtion, the network could not detect any diferences between the two classes. This makes sense because both classes have same tumor type and have same disease. With a pretrained VGG-16 and with a very sharp activation fuction such as Relu, the performance was too low. Although inside the dataset there is a low variety, the performance of 2 frozen block case is much better than the 4 frozen block. This can be explained as follows: Imagenet dataset does not contain any medical image. Pretrained network for medical image analysis would be the best solution for small medical datasets. On the other hand, the pretrained 4 block frozen VGG16 architecture is still much more better than the best conventional machine learning classification combinations.

It can be seen that there is a correlation between best results of machine learning with radiomics features and VGG16 results. Best result tables have many logistic regression classification combinations. Logistic regression classification uses sigmoid function. VGG-16 last layer activation function is also biased sigmoid function. It is clearly seen that sigmoid function can be a good choice for survival time classification problem.

5.4. Survival Analysis

Survival analysis results showed that the degree of radiomics affects hazard ratio differently. Besides, the continuous survival risk predictions over time t is significant in terms of rescheduling therapy or the changing the dosage of therapy. Survival analysis can be used as decision support system for the doctors. A doctor can decide the continuity of a threapy, schedule the therapy day for a specific patient.

It is seen that patients who responded the therapy had less hazard than the other patients. It is seen both in figure 19 and figure 20. Both plots stop around day 1000, because the follow up days are often less than 1000 days. For the patients in the responder class, the risk of death is very low in the first 500 days. After that risk increases. This can be a good indicator that doctors should reschedule the follow up dates for patients in order to early diagnosis or early therapy start.

5.5. Future Work

This thesis can be developed with the following contributions: First of all, there is very few metastatic melanoma patients. Collecting more data effects the deep learning a lot. Besides, other feature selection methods can be implemented for metastatic melanoma studies. For classification tasks, the clinical data can be combined with the hand crafted or deep features. The combination of hand crafted and radiomic features might give good results for survival time classification in metastatic melanoma patients. For Cox regression models, the selected features that come from feature selection algorithms can be used to build a model to investigate survival analysis.

One approach can be done by training a VGG-16 network from strach. Since the dataset has around 22000 images, the result can be improved by training the all layers. But, with only 2 frozen block, the accuracy is 0.92 , and if the frozen blocks are unfrozen, then computational cost would increase a lot. So, fully training approach may increase the accuracy a bit but the computational cost will increase dramatically.

6. Conclusions

This study's first approach is the survival analysis as a classification problem. The main criteria is the days of follow up alive of patients with metastastic melanoma. The patients are divided into two groups according to 1 year survival situations. Second task is to predict therapy response of the patients.
Until now only one paper published radiomic features in metastatic melanoma Durot et al. (2019). Due to high amount of published radiomics papers in other pathology, the state of art articles were collected and their methods were examined. All possible combinations that are the best from published papers were implemented to see the effect of algorithms on metastatic melanoma. Furthermore, this study aimed to see the difference between one slice radiomic features and whole tumor radiomic features. The main contribution of this study is to use a deep neural network to predict the survival time of a patient (less than 1 year or more than one year). Deep neural network is fed by only 2D tumor patches, there is no clinical data. Clinical data is as important as handcrafted or deep features, and even though clinical data is not used, still the results are satisfying. Survival analysis showed that the second order radiomics features have the most significant information for survival time of a patient.

For the results, it can be said that they are sufficient enough and good. They can be improved easily by

adding more data. The main drawback is that the number of patient is only 71 whereas it is seen that state of art publications have 2 time more data than this study. Data is the key for performance in machine learning and artifical intelligence.

7. Acknowledgments

Firstly, I would like to express my gratitude to my supervisors Professor Adrien Bartoli, Dr Benoit Magnin for the fundamental role they have played in this master thesis. I also would like to thank Anne-Flore, who is the student of Dr. Benoit Magnin. Special thanks to Yamid and Kamrul for making my work and my days in Clermont Ferrand easier. I wish all my best to MAIA people, I am so happy to experience this amazing journey with them. Special thanks for Brianna and Mladen for all good days in Girona. I know that our friendship will go long and I will always remember our Girona nights. I would like to thank my colleguages in ENCOV, I spent so nice time with you guys. Most importantly, I want to thank my family. This master helped me to understand that long distance has no importance between family members. I would like to mention and thank to my super Turkish crew; Ogulcan, Laden, Aysegul, Cansever, Han, Caglar, Gonenc, Atakan, Ilknur, Selin, Didem and Firat . Thank you guys, with your support and endless friendships, I was trying to go on this master even in my deepest moments. I could get a chance to start drum lessons in Clermont Ferrand, and special thanks for my teacher Samuel helping me to find a way out of stress. Now my new motto is "If you've got a problem, take it out on a drum".
My most special thank goes to my sister Guler. I am appreciated your support and belief on me all my life. You are the most kind-hearted person I have ever seen and I am so lucky to be your sister. Lastly, I would like to thank all my professors in *MAIA*. Thanks to you, I could get a chance to be part of medicine and help people with computers. This is amazing.
Thanks! - Merci! - Grazie! - Gràcies!- Teşekkurler!

References

Aerts, H.J., Velazquez, E.R., Leijenaar, R.T., Parmar, C., Grossmann, P., Cavalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., Hoebers, F., Rietbergen, M.M., Leemans, C.R., Dekker, A., Quackenbush, J., Gillies, R.J., Lambin, P., 2014. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nature Communications 5. doi:10.1038/ncomms5006.

Afshar, P., Mohammadi, A., Plataniotis, K.N., Oikonomou, A., Benali, H., 2018. From Hand-Crafted to Deep Learning-based Cancer Radiomics: Challenges and Opportunities .

Aydin, L., Kiziltan, E., E, E., Azizaaolu, B., Bekkaraman, A., Doan, S., Ertirk, G., Ku, C., 2017. Is central origin of muscle fatigue distinguished solely in finger tapping performance?, in: 2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE), pp. 542–547. doi:10.1109/BIBE.2017.00009.

Bibault, J.E., Giraud, P., Durdux, C., Taieb, J., Berger, A., Coriat, R., Chaussade, S., Dousset, B., Nordlinger, B., Burgun, A., 2018. Deep Learning and Radiomics predict complete response after neo-adjuvant chemoradiation for locally advanced rectal cancer. Scientific Reports 8, 1–8. doi:10.1038/s41598-018-30657-6.

van der Burgh, H.K., Schmidt, R., Westeneng, H.J., de Reus, M.A., van den Berg, L.H., van den Heuvel, M.P., 2017. Deep learning predictions of survival based on MRI in amyotrophic lateral sclerosis. NeuroImage: Clinical 13, 361–369. doi:10.1016/j.nicl.2016.10.008.

Chaddad, A., Desrosiers, C., Toews, M., Abdulkarim, B., 2017. Predicting survival time of lung cancer patients using radiomic analysis. Oncotarget 8, 104393–104407. doi:10.18632/oncotarget.22251.

D.R.Cox, 1972. Regression Models and Life-Tables. Journal of the Royal Statistical Society. Series B (Methodological) 34, 187–220. doi:10.2307/2985181.

Durot, C., Mulé, S., Soyer, P., Marchal, A., Grange, F., Hoeffel, C., 2019. Metastatic melanoma: pretreatment contrast-enhanced CT texture parameters as predictive biomarkers of survival in patients treated with pembrolizumab. European Radiology doi:10.1007/s00330-018-5933-x.

Gillies, R.J., Kinahan, P.E., Hricak, H., 2016. radiol.2E2015151169. Radiology 278. doi:10.1148/radiol.2015151169.

Haarburger, C., Weitz, P., Rippel, O., Merhof, D., 2018. Image-based Survival Analysis for Lung Cancer Patients using CNNs .

Huang, G.B., Zhu, Q.Y., Siew, C.K., 2006. Extreme learning machine: Theory and applications. Neurocomputing 70, 489 – 501. doi:https://doi.org/10.1016/j.neucom.2005.12.126. neural Networks.

Lacki, M., 2017. Intelligent Prediction of Ship Maneuvering. TransNav, the International Journal on Marine Navigation and Safety of Sea Transportation 10, 511–516. doi:10.12716/1001.10.03.17.

Lao, J., Chen, Y., Li, Z.C., Li, Q., Zhang, J., Liu, J., Zhai, G., 2017. A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme. Scientific Reports 7, 1–8. doi:10.1038/s41598-017-10649-8.

Larkin, J., Ascierto, P.A., Drno, B., Atkinson, V., Liszkay, G., Maio, M., Mandal, M., Demidov, L., Stroyakovskiy, D., Thomas, L., de la Cruz-Merino, L., Dutriaux, C., Garbe, C., Sovak, M.A., Chang, I., Choong, N., Hack, S.P., McArthur, G.A., Ribas, A., 2014. Combined vemurafenib and cobimetinib in BRAF-mutated melanoma. N. Engl. J. Med. 371, 1867–1876. doi:10.1056/NEJMoa1408868.

Leger, S., Zwanenburg, A., Pilz, K., Lohaus, F., Linge, A., Zöphel, K., Kotzerke, J., Schreiber, A., Tinhofer, I., Budach, V., Sak, A., Stuschke, M., Balermpas, P., Rödel, C., Ganswindt, U., Belka, C., Pigorsch, S., Combs, S.E., Mönnich, D., Zips, D., Krause, M., Baumann, M., Troost, E.G., Löck, S., Richter, C., 2017. A comparative study of machine learning methods for time-To-event survival data for radiomics risk modelling. Scientific Reports 7, 1–11. doi:10.1038/s41598-017-13448-3.

Lens, M.B., Dawes, M., 2004. Global perspectives of contemporary epidemiological trends of cutaneous malignant melanoma. British Journal of Dermatology 150, 179–185. doi:10.1111/j.1365-2133.2004.05708.x.

Li, H., Zhong, H., Boimel, P., Ben-Josef, E., Xiao, Y., Fan, Y., 2017. Deep Convolutional Neural Networks for Imaging Based Survival Analysis of Rectal Cancer Patients. International Journal of Radiation Oncology*Biology*Physics 99, S183. doi:10.1016/j.ijrobp.2017.06.458.

LIANG, K.Y., ZEGER, S.L., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73, 13–22. doi:10.1093/biomet/73.1.13.

Lubner, M.G., Smith, A.D., Sandrasegaran, K., Sahani, D.V., Pickhardt, P.J., 2017. CT Texture Analysis: Definitions, Applications, Biologic Correlates, and Challenges. Radiographics 37, 1483–1503. doi:10.1148/rg.2017170056.

Lusa, L., Others, 2013. SMOTE for high-dimensional class-imbalanced data. BMC bioinformatics 14, 106. doi:10.1186/1471-2105-14-106.

Martin-Liberal, J., Kordbacheh, T., Larkin, J., 2015. Safety of pembrolizumab for the treatment of melanoma. Expert Opin Drug Saf 14, 957–964. doi:10.1517/14740338.2015.1021774.

Napel, S., Mu, W., Jardim-Perassi, B.V., Aerts, H.J., Gillies, R.J., 2018. Quantitative imaging of cancer in the postgenomic era: Radio(geno)mics, deep learning, and habitats. Cancer 124, 4633–4649. doi:10.1002/cncr.31630.

Ng, F., Kozarski, R., Ganeshan, B., Goh, V., 2013. Assessment of tumor heterogeneity by CT texture analysis: Can the largest cross-sectional area be used as an alternative to whole tumor analysis? European Journal of Radiology 82, 342–348. doi:10.1016/j.ejrad.2012.10.023.

Nie, D., Lu, J., Zhang, H., Adeli, E., Wang, J., Yu, Z., Liu, L.Y., Wang, Q., Wu, J., Shen, D., 2019. Multi-Channel 3D Deep Feature Learning for Survival Time Prediction of Brain Tumor Patients Using Multi-Modal Neuroimages. Scientific Reports 9, 1–14. doi:10.1038/s41598-018-37387-9.

Nioche, C., Orlhac, F., Boughdad, S., Reuze, S., Goya-Outi, J., Robert, C., Pellot-Barakat, C., Soussan, M., erique Frouin, F., Buvat, I., 2018. Lifex: A freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. Cancer Research 78, 4786–4789. doi:10.1158/0008-5472.CAN-18-0125.

Oakden-Rayner, L., Carneiro, G., Bessen, T., Nascimento, J.C., Bradley, A.P., Palmer, L.J., 2017. Precision Radiology: Predicting longevity using feature engineering and deep learning methods in a radiomics framework. Scientific Reports 7, 1–13. doi:10.1038/s41598-017-01931-w.

Oikonomou, A., Khalvati, F., Tyrrell, P.N., Haider, M.A., Tarique, U., Jimenez-Juan, L., Tjong, M.C., Poon, I., Eilaghi, A., Ehrlich, L., Cheung, P., 2018. Radiomics analysis at PET/CT contributes to prognosis of recurrence and survival in lung cancer treated with stereotactic body radiotherapy. Scientific Reports 8, 1–11. doi:10.1038/s41598-018-22357-y.

Pan, S.J., Yang, Q., 2010. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering 22, 1345–1359. doi:10.1109/TKDE.2009.191.

Paul, R., Hawkins, S.H., Balagurunathan, Y., Schabath, M.B., Gillies, R.J., Hall, L.O., Goldgof, D.B., 2016. Deep Feature Transfer Learning in Combination with Traditional Features Predicts Survival Among Patients with Lung Adenocarcinoma. Tomography 2. doi:10.18383/j.tom.2016.00211.

Ramachandran, P., Zoph, B., Le, Q.V., 2017. Searching for Activation Functions , 1–13.

Rao, S.X., Lambregts, D.M., Schnerr, R.S., Beckers, R.C., Maas, M., Albarello, F., Riedl, R.G., Dejong, C.H., Martens, M.H., Heijnen, L.A., Backes, W.H., Beets, G.L., Zeng, M.S., Beets-Tan, R.G., 2016. CT texture analysis in colorectal liver metastases: A better way than size and volume measurements to assess response to chemotherapy? United European Gastroenterol J 4, 257–263. doi:10.1177/2050640615601603.

Robert, C., Schachter, J., Long, G.V., Arance, A., Grob, J.J., Mortier, L., Daud, A., Carlino, M.S., McNeil, C., Lotem, M., Larkin, J., Lorigan, P., Neyns, B., Blank, C.U., Hamid, O., Mateus, C., Shapira-Frommer, R., Kosh, M., Zhou, H., Ibrahim, N., Ebbinghaus, S., Ribas, A., investigators, K.., 2015. Pembrolizumab versus Ipilimumab in Advanced Melanoma. The New England journal of medicine 372, 2521–32. doi:10.1056/NEJMoa1503093.

Royall, R.M., 1986. Model robust confidence intervals using maximum likelihood estimators. International Statistical Review / Revue Internationale de Statistique 54, 221–226.

Sabaila, A., Fauconnier, A., Huchon, C., 2015. Chimiothérapie intrapéritonéale pressurisée en aérosol (CIPPA): Une nouvelle voie d'administration dans les carcinoses péritonéales d'origine ovarienne. Gynecologie Obstetrique et Fertilite 43, 66–67. doi:10.1002/ijc.29210.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition , 1–14.

Smith, A.D., Gray, M.R., del Campo, S.M., Shlapak, D., Ganeshan, B., Zhang, X., Carson, W.E., 2015. Predicting Overall Survival in Patients With Metastatic Melanoma on Antiangiogenic Therapy and RECIST Stable Disease on Initial Posttherapy Images Using CT Texture Analysis. American Journal of Roentgenology 205, W283–W293. URL: http://www.ajronline.org/doi/10.2214/AJR.15.14315, doi:10.2214/AJR.15.14315.

Sun, W., Jiang, M., Dang, J., Chang, P., Yin, F.F., 2018. Effect of machine learning methods on predicting NSCLC overall survival time based on Radiomics analysis. Radiation Oncology 13, 1–8. doi:10.1186/s13014-018-1140-9.

Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Jianming Liang, 2016. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? IEEE transactions on medical imaging 35, 1299–1312.

Truhn, D., Schrading, S., Haarburger, C., Schneider, H., Merhof, D., Kuhl, C., 2018. Radiomic versus Convolutional Neural Networks Analysis for Classification of Contrast-enhancing Lesions at Multiparametric Breast MRI. Radiology 290, 290–297. doi:10.1148/radiol.2018181352.

Wu, E., Hadjiiski, L.M., Samala, R.K., Chan, H.p., Cha, K.H., Richter, C., Cohan, R.H., Caoili, E.M., Paramagul, C., Alva, A., Weizer, A.Z., 2019. Deep Learning Approach for Assessment of Bladder Cancer Treatment Response. Tomography 5, 201–208. doi:10.18383/j.tom.2018.00036.

Xiaohong Zhang, 2006. Generalized estimating equations for clustered survival data. Retrospective Theses and Dissertations .

Yin, P., Mao, N., Zhao, C., Wu, J., Sun, C., Chen, L., Hong, N., 2019. Comparison of radiomics machine-learning classifiers and feature selection for differentiation of sacral chordoma and sacral giant cell tumour based on 3D computed tomography features. European Radiology 29, 1841–1847. doi:10.1007/s00330-018-5730-6.

Zaharchuk, G., Gong, E., Wintermark, M., Rubin, D., Langlotz, C.P., 2018. Deep learning in neuroradiology. American Journal of Neuroradiology 39, 1776–1784. doi:10.3174/ajnr.A5543.

Zhang, Y., Oikonomou, A., Wong, A., Haider, M.A., Khalvati, F., 2017. Radiomics-based Prognosis Analysis for Non-Small Cell Lung Cancer. Scientific Reports 7, 1–8. doi:10.1038/srep46349.

Appendix A. Radiomics Features and Machine Learning Classifiers Results

lr:logistic regression, rf: random forest , svm: support vector machine, knn:k-nearest neighbour, 3D: 3D radiomics features, 2D: 2D radiomics features

Type	Classification Task	Feature Selection+Classification	Smote (Y/N)	Acc	Sensitivity	Specifity
2D	Survival time prediction	boruta+knn+nosmote	N	0.64	1.0	0.76
2D	Survival time prediction	boruta+lr	N	0.66	0.7215	0.6817
2D	Survival time prediction	boruta+svm	Y	0.7042	0.6222	0.8846
2D	Survival time prediction	boruta+rf	Y	0.7042	0.6222	0.8846
2D	Survival time prediction	boruta+knn	Y	0.7042	1.0	0.1923
2D	Survival time prediction	boruta+lr	Y	0.66	0.9333	0.1923
2D	Survival time prediction	recursive+knn	N	0.7183	0.9777	0.2692
2D	Survival time prediction	recursive+lr	N	0.7042	1.0	0.1923
2D	Survival time prediction	recursive+rf	N	0.55	0.1176	0.9473
2D	Survival time prediction	recursive+svm	N	0.55	0.1176	0.9473
2D	Survival time prediction	recursive+knn	Y	0.55	0.1176	0.9473
2D	Survival time prediction	recursive+lr	Y	0.6478	0.9032	0.1153
2D	Survival time prediction	recursive+rf	Y	0.6338	0.9333	0.1153
2D	Survival time prediction	recursive+svm	Y	0.7605	0.9333	0.4615
2D	Survival time prediction	relief+knn	N	0.7605	0.9333	0.4615
2D	Survival time prediction	relief+lr	N	0.6338	0.9333	0.1153
2D	Survival time prediction	relief+rf	N	0.6338	0.9333	0.1153
2D	Survival time prediction	relief+svm	N	0.6338	0.9333	0.1153
2D	Survival time prediction	relief+knn	Y	0.55	0.1176	0.9473
2D	Survival time prediction	relief+lr	Y	0.7042	0.6222	0.8846
2D	Survival time prediction	relief+rf	Y	0.55	0.1176	0.9473
2D	Survival time prediction	relief+svm	Y	0.7042	0.6222	0.8846
2D	Survival time prediction	rf+knn	N	0.6056	0.9072	0.1735
2D	Survival time prediction	rf+lr	N	0.6447	0.9375	0.1428
2D	Survival time prediction	rf+rf	N	0.6479	0.9213	0.4718
2D	Survival time prediction	rf+svm	N	0.6447	0.9375	0.1428
2D	Survival time prediction	rf+knn	Y	0.6447	0.9375	0.1428
2D	Survival time prediction	rf+lr	Y	0.6338	0.9333	0.1153
2D	Survival time prediction	rf+rf	Y	0.7605	0.9333	0.4615
2D	Survival time prediction	rf+svm	Y	0.6447	0.9375	0.1428
2D	Survival time prediction	sfs+knn	N	0.6478	0.5111	0.8846
2D	Survival time prediction	sfs+lf	N	0.7042	0.6222	0.8846
2D	Survival time prediction	sfs+svm	N	0.6447	0.9375	0.1428
2D	Survival time prediction	sfs+rf	N	0.6478	0.5111	0.8846
2D	Survival time prediction	sfs+knn	Y	0.55	0.1176	0.9473
2D	Survival time prediction	sfs+lr	Y	0.7183	0.9777	0.2692
2D	Survival time prediction	sfs+svm	Y	0.6447	0.9375	0.1428
2D	Survival time prediction	sfs+rf	Y	0.7273	0.9677	0.2992
3D	Survival time prediction	boruta+rf	N	0.6619	0.9333	0.1923
3D	Survival time prediction	boruta+rf	Y	0.7183	0.9777	0.2692
3D	Survival time prediction	boruta+knn	N	0.6338	0.9333	0.1153
3D	Survival time prediction	boruta+lr	N	0.6617	1.0	0.76
3D	Survival time prediction	boruta+svm	N	0.6478	1.0	0.03
3D	Survival time prediction	boruta+knn	Y	0.7183	0.9777	0.2692
3D	Survival time prediction	boruta+lr	Y	0.6315	0.9375	0.1077
3D	Survival time prediction	boruta+svm	Y	0.6447	0.9375	0.1428
3D	Survival time prediction	recursive+rf	Y	0.7606	0.6444	0.9615
3D	Survival time prediction	recursive+rf	N	0.7323	0.6444	0.9614
3D	Survival time prediction	recursive+knn	N	0.7883	0.7111	0.9230
3D	Survival time prediction	recursive+lr	N	0.8169	0.8881	0.6923

Type	Classification Task	Feature Selection+Classification	Smote (Y/N)	Acc	Sensitivity	Specifity
3D	Survival time prediction	recursive+svm	N	0.7323	0.9333	0.3846
3D	Survival time prediction	recursive+knn	Y	0.7887	0.6666	1.0
3D	Survival time prediction	recursive+lr	Y	0.8591	0.9111	0.7692
3D	Survival time prediction	recursive+svm	Y	0.7605	0.9333	0.4615
3D	Survival time prediction	relief+rf	N	0.7323	0.6222	0.9230
3D	Survival time prediction	relief+rf	Y	0.7042	0.6222	0.8846
3D	Survival time prediction	relief+svm	N	0.7042	1.0	0.1923
3D	Survival time prediction	relief+lr	N	0.7042	1.0	0.1923
3D	Survival time prediction	relief+knn	N	0.7887	0.7333	0.8846
3D	Survival time prediction	relief+knn	Y	0.8169	0.7555	0.9230
3D	Survival time prediction	relief+lr	Y	0.7183	1.0	0.23
3D	Survival time prediction	relief+svm	Y	0.6901	1.0	0.1538
3D	Survival time prediction	rf+knn	N	0.7605	0.68	0.8846
3D	Survival time prediction	rf+knn	Y	0.7042	0.6222	0.8461
3D	Survival time prediction	rf+lr	N	0.8169	0.8444	0.7692
3D	Survival time prediction	rf+lr	Y	0.8028	0.8881	0.8076
3D	Survival time prediction	rf+rf	N	0.6760	0.5555	0.8846
3D	Survival time prediction	rf+rf	Y	0.6478	0.5111	0.8846
3D	Survival time prediction	rf+svm	N	0.7323	1.0	0.2692
3D	Survival time prediction	rf+svm	Y	0.7183	0.9777	0.2692
3D	Survival time prediction	sfs+knn	N	0.7183	0.6888	0.7692
3D	Survival time prediction	sfs+knn	Y	0.7042	0.5777	0.9230
3D	Survival time prediction	sfs+lr	N	0.8169	0.9777	0.5384
3D	Survival time prediction	sfs+lr	Y	0.8169	0.9777	0.5384
3D	Survival time prediction	sfs+rf	N	0.6901	0.5333	0.9615
3D	Survival time prediction	sfs+rf	Y	0.5492	0.3111	0.9615
3D	Survival time prediction	sfs+svm	Y	0.7183	1.0	0.2307
3D	Survival time prediction	sfs+svm	N	0.7746	1.0	0.3846
2D	Therapy response	boruta+knn	N	0.7083	0.6764	0.7368
2D	Therapy response	boruta+lr	N	0.5833	0.1470	0.9736
2D	Therapy response	boruta+svm	Y	0.6944	0.6470	0.7368
2D	Therapy response	boruta+svm	N	0.7083	0.6176	0.7894
2D	Therapy response	boruta+rf	Y	0.7916	0.9111	0.6842
2D	Therapy response	boruta+rf	N	0.7361	0.7352	0.7368
2D	Therapy response	boruta+knn	Y	0.75	0.7941	0.7105
2D	Therapy response	boruta+lr	Y	0.5694	0.1176	0.9736
2D	Therapy response	recursive+knn	N	0.72	0.5294	0.8947
2D	Therapy response	recursive+lr	N	0.6944	0.5294	0.8421
2D	Therapy response	recursive+rf	N	0.6666	0.6764	0.6578
2D	Therapy response	recursive+svm	N	0.5138	0.0	0.9736
2D	Therapy response	recursive+knn	Y	0.75	0.6470	0.8421
2D	Therapy response	recursive+lr	Y	0.6944	0.6470	0.7368
2D	Therapy response	recursive+rf	Y	0.6666	0.7647	0.5789
2D	Therapy response	recursive+svm	Y	0.6944	0.4411	0.9210
2D	Therapy response	relief+knn	N	0.75	0.7647	0.7300
2D	Therapy response	relief+lr	N	0.55	0.1176	0.9473
2D	Therapy response	relief+rf	N	0.6225	0.5588	0.6842
2D	Therapy response	relief+svm	N	0.5416	0.08	0.9473
2D	Therapy response	relief+knn	Y	0.7361	0.7058	0.7631
2D	Therapy response	relief+lr	Y	0.62	0.47	0.7613
2D	Therapy response	relief+rf	Y	0.5694	0.7352	0.42
2D	Therapy response	relief+svm	Y	0.5833	0.3235	0.8157
2D	Therapy response	rf+knn	N	0.7083	0.5588	0.8421
2D	Therapy response	rf+lr	N	0.5277	0.0	1.
2D	Therapy response	rf+rf	N	0.6944	0.7058	0.6821

Type	Classification Task	Feature Selection+Classification	Smote (Y/N)	Acc	Sensitivity	Specifity
2D	Therapy response	rf+svm	N	0.6388	0.3823	0.8684
2D	Therapy response	rf+knn	Y	0.7222	0.6764	0.7631
2D	Therapy response	rf+lr	Y	0.5277	0.0	1.0
2D	Therapy response	rf+rf	Y	0.7361	0.8529	0.6315
2D	Therapy response	rf+svm	Y	0.6805	0.4705	0.8684
2D	Therapy response	sfs+knn	N	0.7083	0.5882	0.8157
2D	Therapy response	sfs+lr	N	0.5277	0.0	1.0
2D	Therapy response	sfs+svm	N	0.5277	0.0	1.0
2D	Therapy response	sfs+rf	N	0.5684	0.6176	0.5263
2D	Therapy response	sfs+knn	Y	0.6944	0.6764	0.7105
2D	Therapy response	sfs+lr	Y	0.5277	0.0	1.0
2D	Therapy response	sfs+svm	Y	0.5	0.0	0.9473
2D	Therapy response	sfs+rf+	Y	0.6666	0.7647	0.5789
3D	Therapy response	boruta+rf	Y	0.8333	0.8235	0.8421
3D	Therapy response	boruta+rf	Y	0.8472	0.9111	0.7894
3D	Therapy response	boruta+knn	Y	0.8472	0.9111	0.7894
3D	Therapy response	boruta+lr	N	0.75	0.5882	0.8947
3D	Therapy response	boruta+svm	N	0.8055	0.6470	0.9473
3D	Therapy response	boruta+knn	N	0.8055	0.7647	0.8421
3D	Therapy response	boruta+lr	Y	0.8194	0.7941	0.8421
3D	Therapy response	boruta+svm	Y	0.8611	0.9411	0.7894
3D	Therapy response	recursive+rf	N	0.7916	0.8823	0.7105
3D	Therapy response	recursive+rf	Y	0.7777	0.7058	0.8421
3D	Therapy response	recursive+knn	Y	0.7911	0.6470	0.9210
3D	Therapy response	recursive+lr	N	0.7916	0.5888	0.9736
3D	Therapy response	recursive+svm	N	0.5416	0.05	0.9736
3D	Therapy response	recursive+knn	Y	0.8194	0.7647	0.8684
3D	Therapy response	recursive+lr	Y	0.7777	0.7058	0.8421
3D	Therapy response	recursive+svm	Y	0.6994	0.4117	0.9473
3D	Therapy response	relief+rf	N	0.8194	0.7058	0.9210
3D	Therapy response	relief+rf	Y	0.7361	0.7941	0.6842
3D	Therapy response	relief+svm	N	0.7083	0.3823	1.0
3D	Therapy response	relief+lr	N	0.7222	0.5588	0.8684
3D	Therapy response	relief+knn	N	0.8055	0.7941	0.8157
3D	Therapy response	relief+knn	Y	0.7777	0.7352	0.8157
3D	Therapy response	relief+lr	Y	0.7638	0.6764	0.8421
3D	Therapy response	relief+svm	Y	0.6805	0.3823	0.9473
3D	Therapy response	rf+knn	N	0.8472	0.7647	0.9210
3D	Therapy response	rf+knn	Y	0.8333	0.7647	0.8947
3D	Therapy response	rf+lr	N	0.5277	0.0	1.0
3D	Therapy response	rf+lr	Y	0.5277	0.0	1.0
3D	Therapy response	rf+rf	N	0.7638	0.7647	0.7631
3D	Therapy response	rf+rf	Y	0.7777	0.8235	0.7368
3D	Therapy response	rf+svm	N	0.7777	0.5882	0.9473
3D	Therapy response	rf+svm	Y	0.8194	0.7647	0.8684
3D	Therapy response	sfs+svm	N	0.5277	0.0	1.0
3D	Therapy response	sfs+svm	Y	0.5277	0.0	1.0
3D	Therapy response	sfs+lr	N	0.5277	0.0	1.0
3D	Therapy response	sfs+lr	Y	0.5277	0.0	1.0
3D	Therapy response	sfs+rf	N	0.6944	0.6470	0.7368
3D	Therapy response	sfs+rf	Y	0.8611	0.9117	0.8157
3D	Therapy response	sfs+knn	Y	0.8472	0.8823	0.8157
3D	Therapy response	sfs+knn	N	0.8194	0.7647	0.8684

Appendix B. VGG16 Networks

1. Batch size: 8
2. Number of epochs per patient: 3
3. Optimizer : Adam
4. Loss function: Binary crossentropy
5. Fully connected classifier block: 2 dense layers have *swish* activation function.

The aim is to use as same as possible one VGG16 network for both classification tasks. Thats why batch size, number of epochs per patient and optimizer are used same for both classsification tasks.

These parameters are same for 4 blocks and 2 blocks frozen VGG-16 networks.

Appendix C. Dataset Information

This section is the summary of the dataset that is used in this thesis:

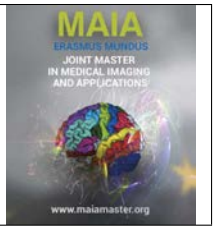
1. Total number of patients: 71
2. Total number of lesions: 539

For therapy response, class 0 has 37 patients and class 1 has 34 patients. Class 0 represents that patients do not response the therapy whereas class 1 shows that patients response the therapy.

For patient survival time classification class 0 has 27 patients and class 1 has 44 patients. The classification based on time between two follow up. It is assumed that if days between two follow up of a patient is more than 1 year, this patient has lived more than one year.

Medical Imaging and Applications

Master Thesis, June 2019



Patch-based segmentation of brain tumor with selective sampling and a U-Net architecture

Liliana Valencia Rodriguez, Mariano Cabezas, Arnau Oliver, Xavier Lladó

Computer Vision and Robotics Group, University of Girona, Catalonia, Spain

Abstract

Gliomas are the most frequent tumors of the central nervous system (CNS) and one of the most deadliest human cancers. The diagnosis is assessed using MRI images, which is a powerful tool to improve the treatment and survival rate of the patients. The multimodal Brain Tumor image Segmentation challenge (BraTS) was proposed as a benchmark for brain tumor segmentation. We use their publicly available BraTS'18 dataset, specifically the training set, to develop a method for brain tumor segmentation. The dataset is composed of 285 cases, each one with images of four modalities: native (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (FLAIR), and ground truth. Our method is based on the simplest but most efficient strategies of the state of the art with U-Net architecture, which have been proved to provide satisfactory segmentation results. We analyze the sampling methods and the loss function using five folds cross-validation with the entire training set. In the case of the sampling methods, we study two approaches, uniform and selective. Regarding the loss functions, we train models with cross-entropy loss and dice loss. The obtained dice scores of our approach are 0.63 for the Enhanced Tumor (ET), 0.75 for the Whole Tumor (WT) and 0.64 the for the Tumor Core (TC). Our results show that in the brain tumor segmentation task, the sampling of the data is fundamental to obtain good segmentation results. A selective sampling is therefore required. The success of the segmentation highly relies on the sampling rather than on sophisticated network architectures, while the dice loss used in the training does not has significant difference in the results.

Keywords: MRI, Brain tumor segmentation, Sampling, U-Net

1. Introduction

Gliomas are the most frequent tumors of the central nervous system (CNS). They include a highly diverse group of primary CNS tumors and were traditionally classified according to their microscopic similarities with putative cells of origin along glial precursor cell lineages (Pisapia, 2017). Nowadays, gliomas are classified based on genotype information about the isocitrate dehydrogenase 1 and 2 (IDH1/IDH2) mutation status. The High Grade Glioma (HGG) are considered the tumors with IDH-wildtype and Low Grade Glioma (LGG) the ones with IDH-mutant. Patients with LGG tumor are generally younger and have a better prognosis. In general, gliomas are one of the most deadliest human cancers (Ostrom et al., 2015).

To assess the gliomas diagnosis, the usage of Magnetic Resonance Imaging (MRI) has become very popular in the latest decade. MRI makes the production of different types of tissue contrast possible by varying excitation and repetition times. This makes it an appropriate tool for imaging different structures of interest (Bauer et al., 2013). The automatic segmentation using MRI images could be a powerful tool to improve the treatment and survival rate of the patients. Tumor segmentation is crucial in tasks such as monitoring the tumor development in patient during therapy, tumor volume measurements and surgical or radiotherapy planning. However, the underlying issue is that gliomas do not have a defined shape or location and usually are mixed with healthy tissues. Hence, these basic characteristics cannot be segmented automatically, which represents a significant challenge.

Many studies have been carried out to improve the automatic diagnosis. In fact, The Multimodal Brain

Tumor image Segmentation Challenge (BraTS) was proposed as a benchmark for brain tumor segmentation in association with the Medical Image Computing and Computer Assisted Interventions (MICCAI) conference 2012 - 2013, and it has been active until now. BraTS provides a publicly available dataset of multi-institutional pre-operative mpMRI scans, and encourages the development of new solutions every year. The image modalities provided by the BraTS'18 are native (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (FLAIR) volumes.

Each of the modalities contains different and useful information used to perform the annotation according to Menze et al. (2015), as shown in Figure 1 and described as follows:

- T2 and FLAIR were used for the segmentation of the *edema*. In T2 images the edema region appears brighter. FLAIR is effective to distinguish between the edema region from the Cerebral-Spinal Fluid (CSF) due to the suppression of water molecules in the imaging process (Liu et al., 2014).
- T1c was used to obtain the segmentation of the *enhancing core* by thresholding the intensities and also to segment the *necrotic core* from the low intensity necrotic structures within the enhancing rim visible in T1c.
- T1 was used together with T1c to segment the gross tumor core. From T1 the inhomogenous component of the hyper-intense lesion and hypointense regions were evaluated. The evaluation of the hyper-intensities was done with T1c. The non-enhancing (solid) core was extracted after the subtraction of the other two structures: enhancing core and the necrotic (or fluid-filled) core.

In the first edition of the BraTS challenge, most of the frameworks were based on traditional machine learning approaches. A discriminative probabilistic approach relying on a random forest classifier was the winner (Bauer et al., 2012), and popularized this method for the 2013 edition.

Nowadays, the application of deep neural networks has become a milestone in the challenge. U-Net architecture has the most extended application for brain tumor segmentation (Dong et al., 2017), (Wang et al., 2017). This architecture was applied with the aim to improve the biomedical image segmentation. It uses skip connections to recover the full spatial resolution in the network output being able to segment fine structures successfully (Ronneberger et al., 2015). The power of the architecture was demonstrated in the BraTS'18 challenge where the second place holders (Isensee et al., 2018) showed that a well trained U-Net with minor modifications provided high segmentation results. Indeed, the authors stressed the importance of the training procedure as well as the importance to consider how the network is trained.

The major problem is related to the high class imbalance and the proposed strategies need to be addressed to solve it. The background, which is around 98% of the image, can highly affect the behaviour of the network. Thus, the way the data is handled has to be analyzed. Patch-based segmentation is a helpful strategy to tackle this problem. An input by patches allows more control on the content of the patches to be extracted, and under some conditions provides more positive samples to the network.

Taking into account the described challenges, the aim of this thesis is to study the current state of the art of the most simple but effective brain tumor segmentation methods and to propose strategies to deal with the segmentation problem. We consider as main reference the approach of Isensee et al. (2018). We analyze the pros and cons of the approach, and implement a segmentation algorithm using the data provided in the BraTS'18 challenge training dataset. The dataset is composed by 210 HGG cases and 75 LGG cases for training.

We focus on the development of deep learning segmentation strategies to provide accurate segmentation that can overcome the class imbalance problem, using a U-Net and a patch-based segmentation method. We use five folds cross-validation with the 285 cases of the training set. To evaluate the behavior of the network, we use cross-entropy loss and dice loss functions, comparing them with different sampling strategies that help us defining the best possible algorithm. Regarding the sampling, the patches are generated inside a brain bounding box to solve a memory problem due the size and amount of images, and the class imbalance problems. For the patches extraction we study two approaches, an uniform sampling and a selective sampling. Moreover, we present qualitative and quantitative results for those strategies. For the quantitative results, Dice Score Coefficient (DSC) is used as a main metric because it is considered a reference metric to assess segmentation.

2. State of the art

Before the BraTS challenge was proposed, the main segmentation methods were divided into two categories: *generative* and *discriminative* methods (Yi et al., 2009) (Menze et al., 2010). The *generative* probabilistic models are very good in the generalization of unseen images, as they use spatial tissue distribution and appearance for the classification. For long time they were the state-of-the-art method for some tumor segmentation tasks. On the other hand, the *discriminative* approach directly learn the relation between the image intensities and the segmentation labels. It concentrates on local features that appear relevant for the tumor segmentation task (Wels et al., 2008). Both methods work, although each

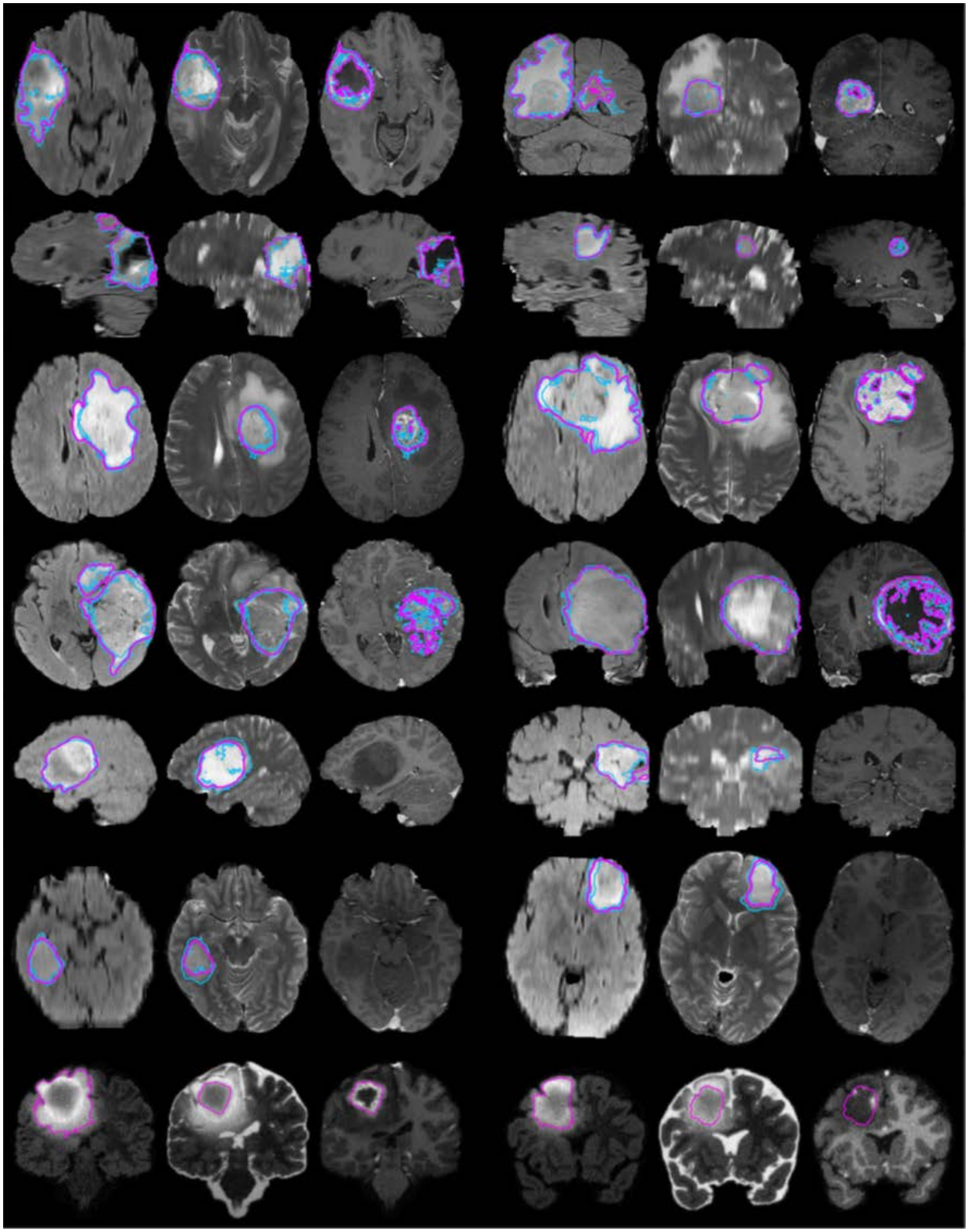


Figure 1: Examples from the BraTS training data, with tumor regions as inferred from the annotations of individual experts (blue lines) and consensus segmentation (magenta lines). Each row shows two cases of high-grade tumor (rows 1−4), low-grade tumor (rows 5−6), or synthetic cases (row 7). Images vary among axial, sagittal, and transversal views, showing for each case: FLAIR with outlines of the whole tumor region (left); T2 with outlines of the core region (center); T1c with outlines of the active tumor region if present (right). Figure taken from Menze et al. (2015).

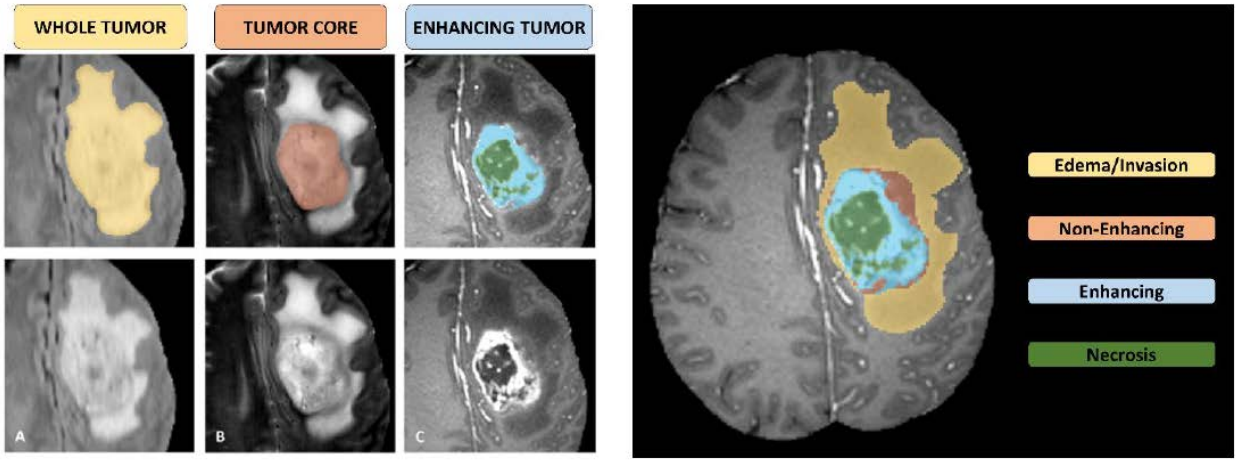


Figure 2: Glioma sub-regions. The image patches show from left to right: the whole tumor (yellow) visible in T2-FLAIR (A), the tumor core (red) visible in T2 (B), the active tumor structures (light blue) visible in T1Gd, surrounding the cystic/necrotic components of the core (green) (C). The segmentations are combined to generate the final labels of the tumor sub-regions (D): ED (yellow), NET (red), NCR cores (green), AT (blue). Figure taken from Menze et al. (2015)

.

has its drawbacks. In order to perform properly, the generative probabilistic models require prior knowledge of the lesion. Atlas-based algorithms which provide spatial information combined with Expectation Maximization (EM) algorithms were widely used for this task (Menze et al., 2010). The tumor is considered a tissue class in the EM algorithm. The main drawback of this strategy is the requirement of accurate registration, which is often difficult. The discriminative models require a massive amount of data in order to overcome the problem of intensity and shape variations, as well as image artifacts. The combination of both models, generative and discriminant has also been considered. For instance, in Tu et al. (2008), a combination of both approaches was used for brain subcortical structure segmentation. The generative model was used to describe the shape while the discriminative approach was used to describe the model appearance.

Regarding the dataset, before BraTS challenge, the developed approaches were tested in private datasets. Every approach was developed using different image protocols and modalities, therefore making an objective comparison between the methods was quite challenging. The BraTS challenge organization proposed a standardized dataset as benchmark for brain tumor segmentation. The methods evaluated with this dataset define the current state of the art. The dataset has been evolving and growing since its creation. The fist dataset, BraTS'12, was composed of 80 cases for training: 30 multi-modal MRI scans (10 LGG cases and 20 HGG cases, with manual annotations), and 50 simulated cases (with same proportion of LGG and HGG). For testing, 16 simulated images (11 HGG cases and 5 LGG cases), and 15 real cases (11 HGG and 4 LGG). It was manually annotated for two tumor labels, edema and core. During the first edition of the challenge, the use of generative and discriminative approaches was crucial. The winner of the challenge applied a discriminative method, reaching average DSC of 0.73 and 0.59 for tumor and edema, respectively (Bauer et al., 2012). In this approach the segmentation task was modeled as a energy minimization problem in a conditional random field. Random forest was used as a classifier and a spatial regularization where the weighting function depends on the voxel spacing in each dimension was applied. During the challenge aftermath, the insufficiency of two tumor classes was considered. The core label contained substructures with very different appearances in the various modalities. Hence, this time, the images were reannotated with four labels divided as stated in Menze et al. (2015):

- Label 1: NCR (Necrotic). Describes the necrotic core which resides in the enhancing rim of the HGG and sometimes appears cystic.
- Label 2: ED (peritumoral edematous/invaded tissue). Describes the peritumoral edematuous and invaded tissue.
- Label 3: NET. describes the non-enhanced areas of the tumor core that in HGG are the surrounding vasogenic edema on T2. In LGG, NET delineates the gross tumor.
- Label 4: AT. Describes the enhancing regions within the gross tumor abnormality, but not the necrotic center.

To evaluate the performance of the algorithms, the organizers of the BraTS challenge use a different configuration of the structure. Three mutually inclusive tumor regions define as Whole Tumor (WT) region, which

Table 1: Number of cases and changes in tasks throughout the years of the BraTS Challenge (Taken from Bakas et al. (2018))

Brats Instance	**Training data**	**Validation data**	**Testing data**	**Tasks**	**Type of data**
2012	35	NA	15	Segmentation	Pre-operative
2013	35	NA	25	Segmentation	Pre-operative
2014	200	NA	38	Segmentation, Assessment of disease progression	Longitudinal
2015	200	NA	53	Segmentation, Assessment of disease progression	Longitudinal
2016	285	NA	191	Segmentation, Assessment of disease progression	Longitudinal
2017	285	46	146	Segmentation Prediction of patient overall survival	Pre-operative
2018	285	66	191	Segmentation Prediction of patient overall survival	Pre-operative

includes all four tumor structures (union of all labels), Tumor Core (TC) region which includes all tumor structures except *edema* (union of labels 1,3 and 4) and Active Tumor (AT) region which only contains *enhancing core* structures unique, which only occurs in HGG cases as shown in Figure 2.

From 2013, the synthetic data was removed from the dataset and the amount of real data has been increasing and expanding the dataset (Table 1 resumes the growth of the dataset). The four labels were kept until 2016. Later, an overestimation of the NET (label 3) by some annotators was noticed and from 2017 on, this label was removed and fused with NCR (label 1). Additionally, contralateral and periventricular regions of T2-FLAIR hyper-intensity were excluded from the ED region, unless they were contiguous with peritumoral ED.

Regarding the methods, and with the popularization of convolutional neural networks (CNN), most of the new approaches were based on these architectures, beating the results of the conventional methods. A simple architecture with 2 convolutional layers that takes as input 2D images (slices) of the axial view, obtained DSC of 0.88, 0.79 and 0.73 for whole tumor, core and enhanced tumor respectively (Havaei et al., 2015). The winner approach from 2015 (Pereira et al., 2016) was based on an individual CNN architecture for each type of glioma. LGG was segmented using a 9 layer architecture and HGG with 11 layers. For the optimization, the Stochastic Gradient Descent (SGD) was used as the main optimizer and the Nesterov's Accelerated Momentum applied in the regions where the curvature was low. Leaky ReLu for the activation function and a normalization proposed by Nyúl and Udupa (1999) used to make contrast and intensity ranges more similar. The algorithm was trained with the BraTS'15 dataset, composed of 220 HGG cases and 54 LGG cases.The DSC score obtained in the challenge, with a testing set of 53 cases of both low and high grade gliomas, was 0.88, 0.83 and 0.77, for whole tumor, core and enhanced tumor respectively. In the posterior challenges, popular 3D CNN architectures started to be employed as DeepMedic (Kamnitsas et al., 2016).

Since BraTS'17, the dataset has bee changed with comparison to the previous version. This dataset contains more clinically-acquired 3T multimodal MRI scans from different institutions and protocols. Moreover, the scans have been categorized as pre or post-operative by expert neuroragiologists. The pre-operative scans were annotated by experts for the different sub-regions following the same annotation protocol. The protocol comprises GD-enhancing tumor (ET label 4), the peritumoral edema (ED label 2), and the necrotic and non-enhancing tumor core (NCR/NET label 1).

On this dataset, the combination of several CNNs also achieved promising results. The approach proposed by Kamnitsas et al. (2017), which was an ensemble of the DeepMedic architecture, Fully Convolutional Network (FCN) and U-Net architecture achieved the best performance in 2017. In their approach, each model was trained individually and provided class-confidence maps as outputs. After ensembling the models, an ensemble's confidence map for each class was computed. Every map was calculated of the average confidence of each model for a voxel. The class assigned to the voxel was the one with the highest confidence. The training was performed with the BraTS'17 dataset which consisted of 210 HGG and 75 LGG cases, and 46 validation cases. The DSC obtained in the challenge, with a test set of 146 cases, was 0.886 for whole tumor, 0.785 for core and 0.729 for enhanced tumor.

The latest dataset, BraTS'18, is composed of 285 cases (210 HGG and 75 LGG cases) for training, 66 cases for validation, and 191 cases for testing. The second place in the challenge was achieved by Isensee et al.

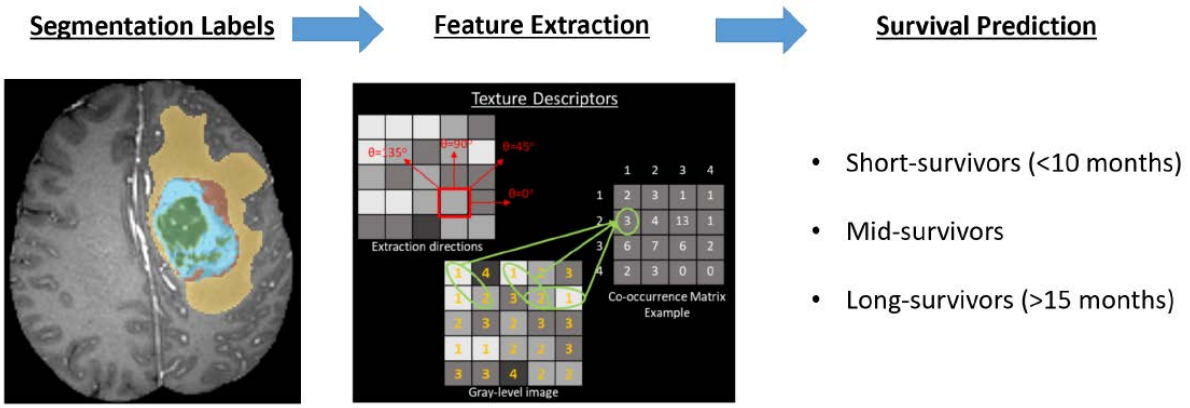


Figure 3: Illustrative pipeline example for predicting patient overall survival. Figure taken from Bakas et al. (2018)

(2018) with an approach focused on the training rather than tuning the network. The DSC score on the validation set was 0.908 for whole tumor, 0.854 for core tumor, and 0.908 for enhancing tumor. They used a 3D U-Net architecture with maxpooling, trilinear upsampling and instance normalization with input patches of 128x128x128. For the activation function, Leaky ReLu was applied, and a multi-class dice loss as a loss function. A co-training private dataset was used to train the model.

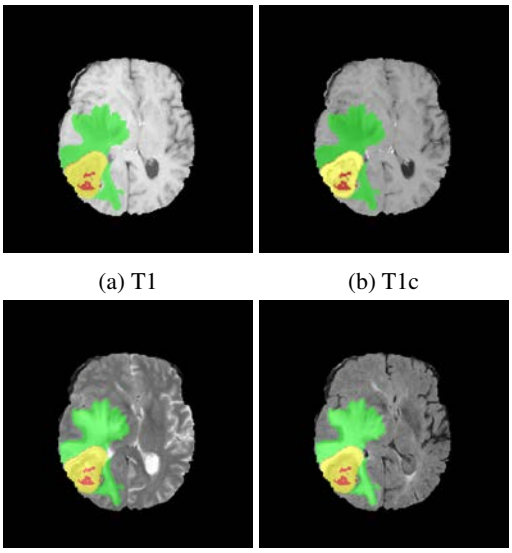
A semantic segmentation of CNNs based on encoder-decoder networks holds the first place in the challenge. In the approach of Myronenko (2018) a larger encoder part is in charge of the image features using the ResNet (He et al., 2016) blocks with normalization and ReLu as activation function. The decoder, dedicated to reconstruct the segmentation mask, has the same structure as the encoder but is smaller, having a single block per spatial level. In the validation set, the DSC obtained was 0.88 for whole tumor, 0.81 for tumor core and 0.76 for enhanced tumor.

It is important to mention that BraTS challenge aims to go beyond the segmentation. The patient survival prediction has been added as a secondary task in the last two editions. For this task, clinical data of patient age, overall survival, and resection status are provided to develop methods that can predict the patient survival via integrative analysis of the radiomic features and machine learning algorithms. The basic pipeline suggested by the BraTS organizer for the survival task is shown in Figure 3.

3. Material and methods

3.1. Dataset

The dataset used for the development of this master thesis is the one provided by the BraTS Challenge 2018. The original characteristics of the BraTS dataset



(a) T1 (b) T1c

(c) T2 (d) FLAIR

Figure 4: Case Brats18_TCIA02_377 from BraTS'18 dataset. The four modalities of the BraTS'18 dataset with segmentation label.

are shown in Table 2. The dataset is highly variate, containing samples from 19 centers and the acquisition protocols are therefore diverse. It is pre-processed with co-registration to the same anatomical template, the SRI24 atlas (Rohlfing et al., 2010), then skull stripped and interpolated to the same resolution (1 mm^3) (Bakas et al., 2017). It contains 285 cases for training: 210 HGG and 75 LGG, which are the ones used for this implementation. All the images have the same size of 240x240x155. In Figure 4 different modalities with segmentation labels exemplify the dataset.

3.2. Methods

Our proposed method for the brain tumor segmentation includes pre-processing of the images, patch extraction using different sampling approaches, training a generic 3D U-Net model with two different loss functions and segmentation, as shown in Figure 5. U-Net

Table 2: Original characteristics of the BraTS dataset (Taken from Bakas et al. (2018)).

Acronym	MRI sequence	Property	Acquisition	Slice thickness
T1	T1-weighted	Native image	Sagittal or Axial	Variable (1-5mm)
T1Gd	T1-weighted	Post-contrast enhancement (Gadolinium)	Axial 3D acquisition	Variable
T2	T2-weighted	Native image	Axial 2D	Variable (2-4mm)
T2-FLAIR	T2-weighted	Native image	Axial or Coronal or Sagittal 2D	Variable

architecture consists on a contracting path which captures context and a symmetric expanding path that enables precise location (Ronneberger et al., 2015). This architecture was built based on the Fully convolutional network (Long et al., 2015), where the usual contracting network is replaced by successive layers with upsampling operations, increasing the resolution of the output. Furthermore, high resolution features from each path, contracting and upsampled output, are combined to improve the localization accuracy. Since the label assignment has to be done per voxel, preserving the localization is one of the keys of this network.

The U-Net architecture implemented is based on the proposal of Isensee et al. (2018) with some modifications, such as reduction in the features channels at the highest resolution and the size of the patches (Figure 7). The modifications consider the limitation in the resources that we have in terms of GPU memory. Also, we believe that the proposed sampling strategies increase the number of patches and help the learning process.

3.2.1. Pre-processing

The images are normalized to a normal distribution by subtracting the mean and dividing by the standard deviation for each modality. To reduce the background voxels and the training time, a bounding box of the brain is extracted from the image. Thus, just the pixels with intensity higher than zero are kept. The patches are extracted inside the brain bounding box. Some shape problem may arise during the patch extraction, such as patches lying in the borders of the image, containing less voxels than the indicated patch size. Consequently, the images are zero padded to the size of the patch in each dimension.

3.2.2. Patch sampling

The sampling of the data is also crucial in the learning process. We defined a uniform sampling, similar to a sliding window, where the input data are patches that overlap. There is no distinction between classes and the full image, in patches, will go through the network. To do so, a list of centers that defines the patches is created. This list will define the location of the patches. The patches shape are 32x32x32 with steps of 16x16x16.

A second sampling called selective sampling is proposed. We take advantage of the prior knowledge of the ground truth image to choose the centers that define the patches for the labels 1,2 and 4. For label 0, which is the background, we use the uniform sampling. An increased number of positive samples will result from this strategy and therefore a better prediction of each label is expected. The total amount of patches will directly relate to the amount of voxels with tumor's label in each volume. The size of the patches is 32x32x32. In the case of label 0 the step is 16x16x16.

3.2.3. Network architecture

The input of the network are patches of 32x32x32 and batch size of 150. We use 16 feature channels at the highest resolution, resulting in a network with 2.642.980 parameters. The channels in last block of the down path are scale 16 times. In the upsampling path, the network accounts for the size of the blocks to guarantee that the output size is the same that the input. We use ReLu as activation function, pooling of 2x2x2, and trilinear upsampling of 2x2x2. The architecture is shown in Figure 7.

3.2.4. Training

Giving that the background, that is all but tumor, occupies the major area of the image, the learning process often gets trapped into the local minima of the loss function. Therefore, the predictions are biased towards the background, and the target region is often missed or partially detected. The proposed strategy to overcome this problem is to reduce the background presence generating patches from the bounding box area. Moreover, we tested two loss functions, considering the problem in the learning process that generates the high class imbalance of the dataset. The first loss function is the cross-entropy loss function implemented in the Pytorch library. The second one is the dice loss function. In Milletari et al. (2016) a binary dice loss was proposed to overcome the imbalance problem and Drozdzal et al. (2016) proposed a multi-class adaptation that is used our approach and defined as:

$$\mathcal{L}_{dc} = -\frac{2}{|K|}\sum_{k\in K}\frac{\sum_i u_i^k v_i^k}{\sum_i u_i^k + \sum_i v_i^k} \quad (1)$$

where $k \in K$ are the number of classes, u is the softmax output of the network, v is one hot enconding of the ground truth segmentation map, and i the number of pixels in the training patch.

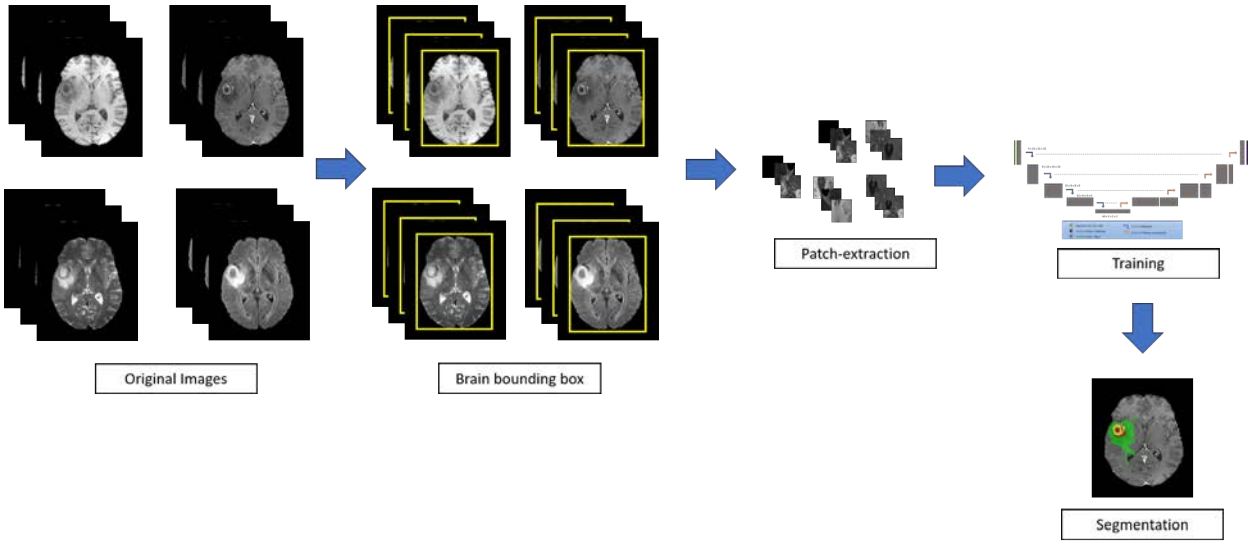


Figure 5: Pipeline for the brain tumor segmentation strategy implemented defining the tasks: brain bounding box patch extraction, training and segmentation.

3.2.5. *Evaluation*

Given that we want to compare our results with the architecture proposed by Isensee et al. (2018), we use the same evaluation with five folds cross-validation of the 285 cases of the training set for our different configurations. We use 10 epochs and early stopping criteria where the loss is evaluated with respect to the best loss result. We define a patient of 3 epochs.

At the testing stage, we perform the segmentation of the full patience at once. In this way, we avoid the problems that may arise in the reconstruction if the testing is done by patches.

The evaluation is obtained from the online evaluation platform of the BraTS challenge website for the training dataset. The results provided by the online evaluation comprise of three labels: whole tumor (WT), tumor core (TC) and enhanced tumor (ET). The metrics are DSC, sensitivity, specificity and Hausdorff distance.

DSC is defined as:

$$DSC = \frac{2\,|X \bigcap Y|}{|X| + |Y|} \tag{2}$$

where X is the ground truth and Y the resulting segmentation.

The sensitivity or True Positive Rate (TPR) (3) and specificity or True Negative Rate (TNR) (4) of each class is computed as:

$$TPR = \frac{TP}{TP + FN} \tag{3}$$

$$TNR = \frac{TN}{TN + FP} \tag{4}$$

where TP are the true positive voxels, TN the true negative voxels and FN the false negative voxels.

The Hausdorff distance measures how far are the two subset of regions, defined by the classes, from each other. Hence, for all the points p on the surface ∂P_1 of the given value P_1 the shortest least-square distance (p, t) is calculated to points t on the surface ∂T_1 of the other given volume T_1 and vice versa, returning the maximum value overall d

$$Haus(P,T) = max\left\{ \sup_{p\in\partial P_1} \inf_{t\in\partial T_1} d(p,t), \sup_{p\in\partial T_1} \inf_{t\in\partial P_1} d(p,t) \right\} \tag{5}$$

Since we have to defined what is the best configuration, we use paired t-test with the mean dices of each label. The paired t-test is performed for samples that are connected. In our case, we evaluate the dice for each label in each patient changing the sampling method or loss function in the configuration. The resulting P-value

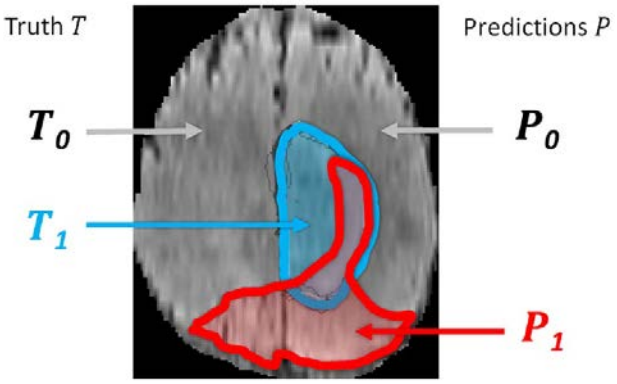


Figure 6: Region T_1 is the true lesion area (outline blue),T_0 is the remaining normal area. P_1 is the area that is predicted to be lesion byfor examplean algorithm (outlined red), and P_0 is predicted to be normal. Figure taken from Menze et al. (2015)

.

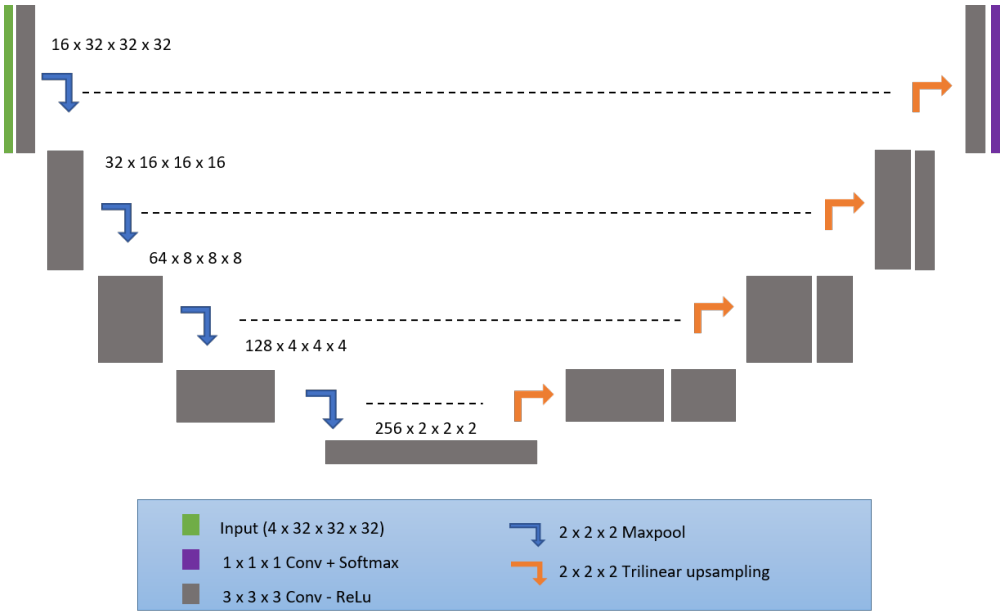


Figure 7: U-Net architecture with some modifications with respect to Isensee et al. (2018). ReLu as activation function, patch size and number of filters

.

from the paired t-test tell us if there is statistically significant difference in DSC among the experiments considering label-by-label approach. We defined the significance level $alpha = 0.01$. For ET label, we did not consider the cases where label 4 is not present in the ground truth. P-Values can also be interpreted as the probability that the results occurred by chance.

3.2.6. Implementation

The deep learning framework used for implementation is Pytorch in Python 3.6. We use a GPU GeForce GTX 1080 with 12 GB of memory to perform the experiments.

4. Results

To evaluate the impact of the sampling method in the resulting model, we train a model for each method of sampling in a five folds cross-validation configuration of the 285 cases. Dice loss function is used during the training, obtaining mean DSC results for uniform sampling of 0.483 for ET, 0.543 for WT and 0.540 for TC. For the selective sampling we obtained 0.629 for ET, 0.755 for WT and 0.708 for TC. DSC are shown in figure 8. For other quantitative results, see Table 3.

Then, we perform a paired t-test for each class obtaining values of $p > 0.01$ for all the labels which shows that they are statistically different between methods of sampling.

We use the previous obtained model, train with dice loss and selective sampling as base line. To compare if the loss function affect the resulting model, we train

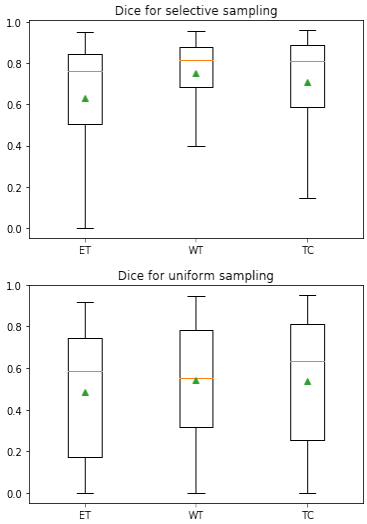


Figure 8: Dice scores for each label, ET, WT and TC resulting from the segmentation model trained for each sampling method in five folds cross-validation of 285 cases. The green triangle represent the mean value for each class. In the selective sampling the mean DSC are: 0.629 for ET, 0.755 for WT and 0.708 for TC. In the uniform sampling the mean DSC are: 0.483 for ET, 0.5428 for WT and 0.5400 for TC.

a model with cross-entropy loss function and selective sampling. The DSC results obtained are 0.594 for ET, 0.756 for WT and 0.727 for TC (Figure 9). For other metrics, refer to Table 4.

The paired t-test for the loss comparison shows that the models trained with cross-entropy loss and dice loss are not statistically different with $p < 0.01$.

In table 5 qualitative results of the experiments are shown. The first two rows show the cases where the seg-

Table 3: Quantitative results of segmentation for models generated with uniform and selective sampling respectively, in five folds cross-validation of 285 case. Metrics were computed by the online evaluation platform.

	Mean DSC		Sensitivity		Specificity		Hausdorff distance	
Label	**Uniform**	**Selective**	**Uniform**	**Selective**	**Uniform**	**Selective**	**Uniform**	**Selective**
ET	0.483 ± 0.307	0.629 ± 0.304	68%	75%	98%	0.99%	44.351	23.456
WT	0.543 ± 0.255	0.755 ± 0.168	85%	87%	88%	97%	58.018	51.959
TC	0.540 ± 0.302	0.708 ± 0.243	64%	77%	97%	99%	49.191	40.802

Table 4: Quantitative results of segmentation for models generated with dice loss and cross-entropy loss function respectively, in five folds cross-validation of 285 cases. Metrics were computed by the online evaluation platform.

	Mean DSC		Sensitivity		Specificity		Hausdorff distance	
Label	**DSC**	**CE**	**DSC**	**CE**	**DSC**	**CE**	**DSC**	**CE**
ET	0.629 ± 0.304	0.594 ± 0.303	75%	83%	99%	99%	23.456	29.656
WT	0.755 ± 0.168	0.756 ± 0.166	87%	93%	97%	96%	51.959	47.713
TC	0.708 ± 0.243	0.727 ± 0.227	77%	81%	99%	99%	40.802	38.577

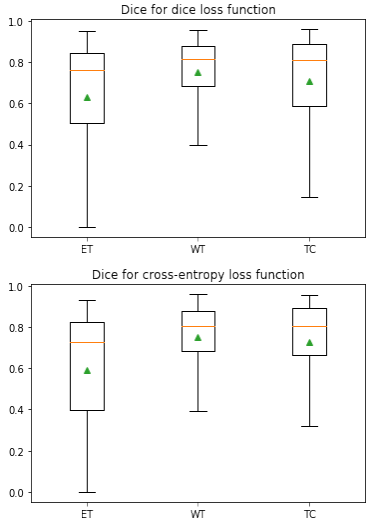


Figure 9: Dice scores for each label, ET, WT and TC resulting from the segmentation model trained for each loss function, dice loss and cross-entropy loss, in five folds cross-validation of 285 cases. The green triangle represent the mean value for each class. In the model trained with dice loss mean DSC are: 0.629 for ET, 0.755 for WT and 0.708 for TC. In the model trained with cross-entropy loss the mean DSC are: 0.594 for ET,0.756 for WT and 0.727 for TC.

mentation results highly agrees with the ground truth. In the last row, segmentation results of one of the most challenge cases is shown. In this case, many slices of the intensity images do not have tumor content, and therefore the prediction ability of the model is tested.

In order to compare our results with the state of the art, we report the DSC results obtained for Isensee et al. (2018) in five folds cross-validation with the 285 cases of the training set without cotraining data. They are: 0.734 for ET, 0.897 for WT and 0.821 for TC.

5. Discussion

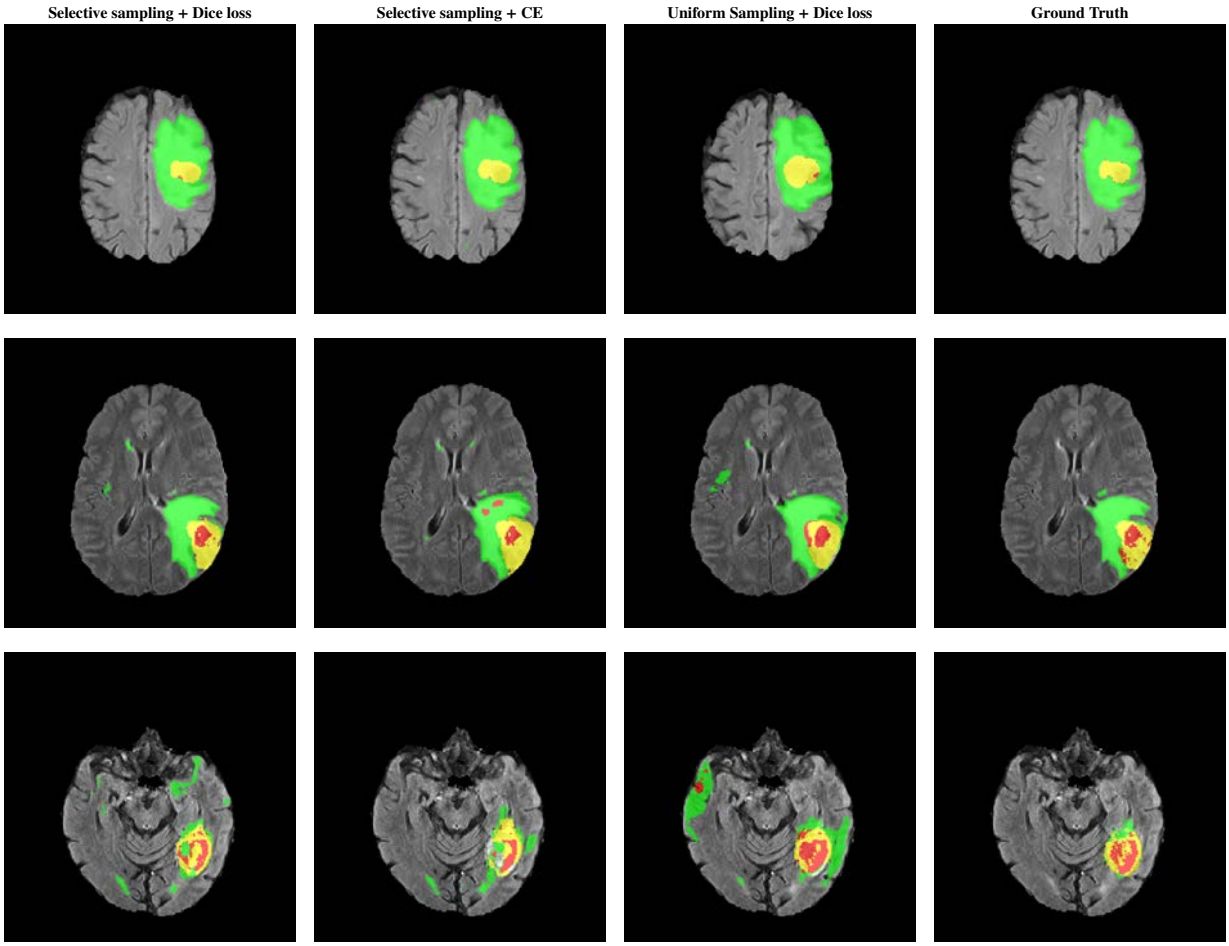
In this master thesis, we analyze the effect of the sampling of the data and the loss function in the generation of models for brain tumor segmentation with U-Net architecture. We study the impact of the sampling, testing an uniform and selective patch sampling. The selective sampling is based on the prior knowledge given in the ground truth. Once we defined the sampling strategy, we analyze the effects of the two loss functions, cross-entropy loss and dice loss, in the training model with different. We finally contrast our results with the state of the art.

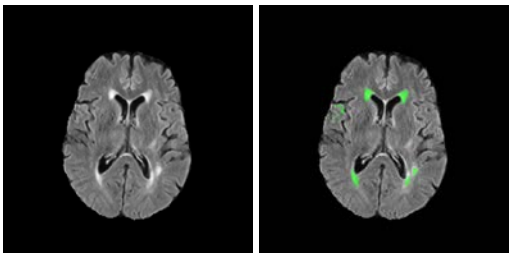
Handling the background is one of the main challenges in the brain tumor. Selecting the proper way to input the data to the network is fundamental to overcome this problem. We test the U-Net architecture with the two defined ways of sampling, uniform and selective, using in both cases dice loss function in the training. In this way, we aim to determine if the resulting model would perform better or worse depending of the sampling. Later, we segment the patients with each model configuration, and send them to the online platform for the training evaluation (Table 3). Qualitative results are shown in Figure 11. For some cases, the DSC for the ET label was zero. This is because this region is not present in all the images of the data set, and at the segmentation time, this label was the most challenging.

Using both sampling methods, we observe that label 2 ED from the training set, is the most miss-classified label out of the tumor region in both sampling methods. This can be due to the intensity similarity with the true voxels belonging to the class. However, in the case of the selective sampling is possible to define a pattern. The hyper-intense voxels out of the tumor region are the ones assigned to this label as shown in Figure 10.

In Figure 8 is possible to visually check that the variability of the dices is large in the uniform sampling

Table 5: Qualitative results. SS refers to the selective sampling, US to the Uniform Sampling and CE to the cross entropy loss function. First row: case Brats18_TCIA04_192. Second row: case Brats18_CBICA_AAP. Last row: case Brats18_CBICA_AQA. The labels are described as: NCR & NET (red), ED (green), ET (yellow).





(a) Flair (b) Segmentation mask

Figure 10: Left: FLAIR image. Right: FLAIR with segmentation mask. Relation between the miss-classified voxels of label 2 with the hyper-intense pixels.

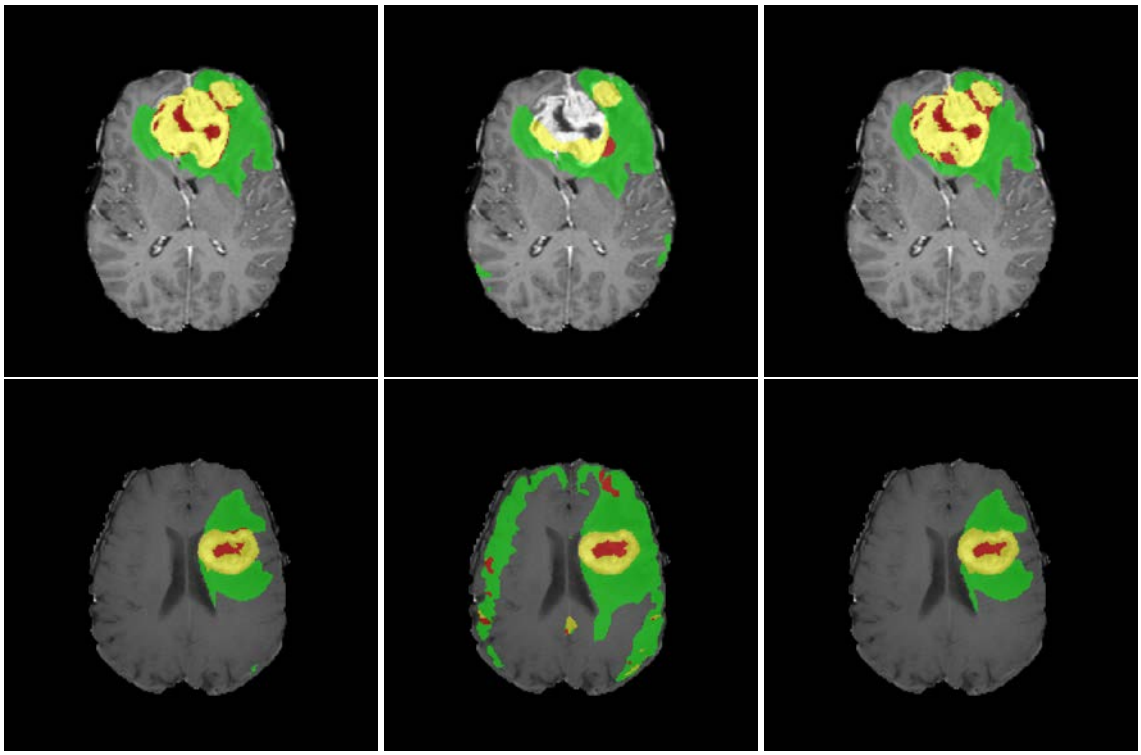
model, also reflected in the standard deviation of the mean. It shows, once again that the model is not able to learn effectively the labels. This is reflected also in the sensitivity results. The model under performs in the classification of the positive samples. For the selective sampling, the mentioned variability, specially in the WT label, improves while there is still room for improvement in the ability of the model to correctly classify the positive samples.

The selective sampling method is chosen considering the statistically significant difference resulting from the paired t-test and the described results. In this way, we conclude that the sampling method highly impacts the resulting model and therefore the segmentation results.

Our hypothesis for the difference between the resulting models with the different sampling methods, is that there is a high probability that the first input samples to the net are background in the case of the uniform sampling. The amount of patches with tumor content is small when compared with the background samples. Then the network could not receive enough inputs to learn them, making the loss function to be completely biased towards the negative sample.

Considering our results, the following trained models use selective sampling.

Regarding the loss function, we were expecting to have statistically significant difference between the cross-entropy loss function and dice loss. Our assumption is done on the basis that the dice loss works better in high class imbalance problems. Its objective is to max-



(a) (b) (c)

Figure 11: Up: images of the case Brats18_2013_17. Down: images of the case Brats18_TCIA01_231. Segmentation masks results for the models trained with uniform and selective sampling where 11a are obtained with the model trained with selective sampling, 11b are obtained with the model trained with uniform sampling and 11c are the ground truths. The labels are described as: NCR & NET (red), ED (green), ET (yellow).

imize the overlap between the prediction and ground truth class, and then it should optimize all the classes, while the cross-entropy loss looks for a general agreement of the classes. Then, once the function identifies correctly one of the classes, it tends to follow the traced path, leaving behind the others. However, our results show that there is not statistical significant difference between them. Indeed, the behaviour in the general DSC distribution is very similar, which is reflected in the standard deviation values. An example of the resulting segmentation if shown in Figure 12.

The model trained with cross-entropy loss also miss-classified hyper-intense voxels. There are some very challenging cases, where both models miss-classified big areas of the brain region (Figure 13). For instance in the case Brats18_TCIA03_133, with segmentation mask shown in Figure 13. In some intermediate slices of the volume, the tumor content is not present . Thus, there is not available reference for the learning process and this represent a challenge in the method development. Our models, in most those slices, assigned labels incorrectly.

Referring to the Hausdorff measure, this metric is very susceptible to small outlying subregions. In the box plot graph of all the approaches, we can observe that we have large outliers represented by the extremes. All the Hausdorff measures where computed over masks that does not have a post-processing step and large Hausdorff distances are therefore expected.

We try to apply the same architecture configuration of Isensee et al. (2018) changing the normalization method between layers. In their strategy, it is not possible to use batch normalization, since they have a very small batch of 2. Hence, the variance will be very close to zero, making the estimations very noisy and impacting the training negatively. In our implementation, we have large batch sizes and we can avoid this problem. However, the obtained results are far from our expectations and our base model. The training time is also considerable high, being approximately 90 minutes per epoch. Thus, we decide to continue with the simpler version of the architecture.

In terms of comparison with the our state-of-the-art reference, our DSC are below Isensee et al. (2018) results in the training set. Nevertheless, we highlight the work developed for the sampling, being our main target for the segmentation task in this master thesis.

6. Conclusions

In our approach, we demonstrate that the most important consideration to have a good segmentation method using deep learning is taking care of the data. By studying two sampling approaches, uniform and selective, we could prove that a generic U-Net configuration can

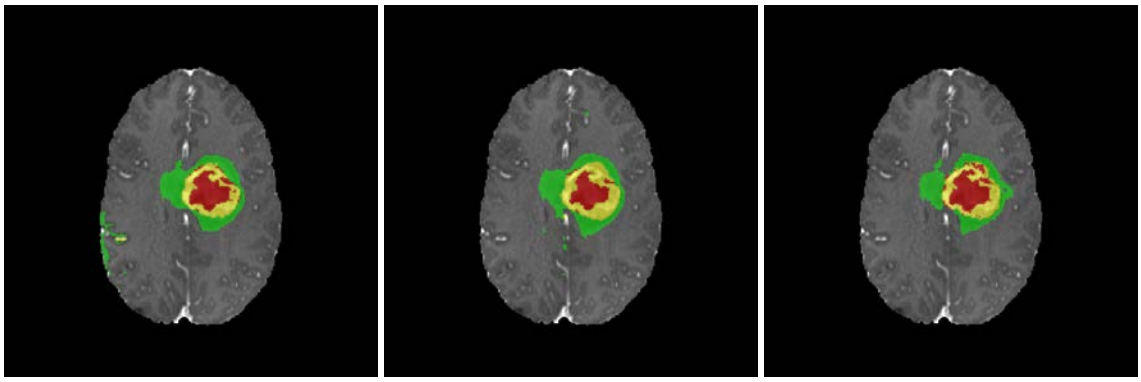


Figure 12: Case Brats18_CBICA_AOO. Left: segmentation masks results for the models trained with dice loss. Center: segmentation mask obtained with the model trained with cross-entropy loss function. Right: ground truth

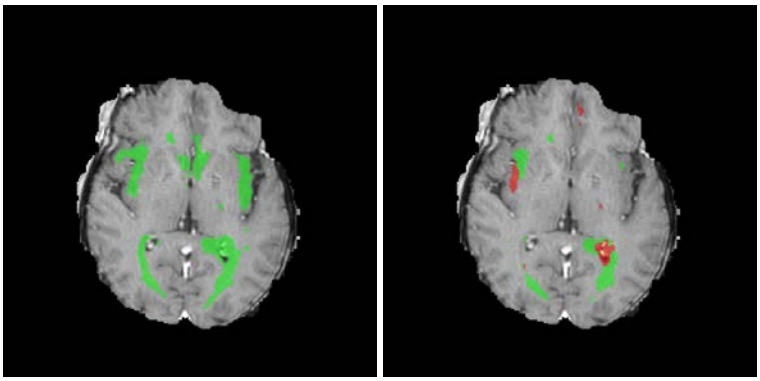


Figure 13: Case Brats18_TCIA03_133. Left: Mask generated for model trained with cross-entropy loss. Right: Mask generated for model trained with cross-entropy loss. The labels are described as: NCR & NET (red), ED (green), ET (yellow)

perform very well when the data is carefully sampled, which is the case of the selective sampling. Regarding the loss function, we could prove that there is not significant difference between the cross-entropy loss function and the dice loss function. Even when they differs in DSC results, both are highly suitable for the brain tumor segmentation tasks. However,it is certainly easier to track the training process with the dice loss because it allows to have an idea of the resulting segmentation. As we mentioned, we do not include a post-processing step which highly affects the Hausdorff distance. Since our main metric was DSC, we preferred to prioritize the results improvements in terms of this metric. This is a task to work on and improve the results of this specific metric. However, it requires an improvement of our approach because some cases have large areas of the brain miss-classified, making impossible to implement a simple solution based on area exclusion. Finally, even when we could not achieve the state-of-the-art results, we consider that our work provides an appropriate base on the sampling strategies.

7. Acknowledgments

I would like to thank my professors and supervisors Dr Xavier Lladó and Dr Arnau Olivier for their support and the opportunity to develop my master thesis in the VICOROB lab. My gratitude to Dr Mariano Cabezas for his patience, support and advice. To Albert Cleriges for his advise, coding lesson and snippets of code. To Dr Sergi Valverde for your inspiration, motivation and advise. To MAIA administration staff, Aina Roldan for her understanding and support.To my MAIA friends who made this experience unique and unforgettable. During challenging days, we stayed together creating links for life time. To my family, my biggest treasure.

References

Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C., 2017. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data 4, 170117.

Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., et al., 2018. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. arXiv preprint arXiv:1811.02629 .

Bauer, S., Fejes, T., Slotboom, J., Wiest, R., Nolte, L.P., Reyes, M., 2012. Segmentation of brain tumor images based on integrated hierarchical classification and regularization, in: MICCAI BraTS Workshop. Nice: Miccai Society.

Bauer, S., Wiest, R., Nolte, L.P., Reyes, M., 2013. A survey of mri-

based medical image analysis for brain tumor studies. Physics in Medicine & Biology 58, R97.

Dong, H., Yang, G., Liu, F., Mo, Y., Guo, Y., 2017. Automatic brain tumor detection and segmentation using u-net based fully convolutional networks, in: annual conference on medical image understanding and analysis, Springer. pp. 506–517.

Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., Pal, C., 2016. The importance of skip connections in biomedical image segmentation, in: Deep Learning and Data Labeling for Medical Applications. Springer, pp. 179–187.

Havaei, M., Dutil, F., Pal, C., Larochelle, H., Jodoin, P.M., 2015. A convolutional neural network approach to brain tumor segmentation, in: BrainLes 2015, Springer. pp. 195–208.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., Maier-Hein, K.H., 2018. No new-net, in: International MICCAI Brainlesion Workshop, Springer. pp. 234–244.

Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., Rajchl, M., Lee, M., Kainz, B., Rueckert, D., et al., 2017. Ensembles of multiple models and architectures for robust brain tumour segmentation, in: International MICCAI Brainlesion Workshop, Springer. pp. 450–462.

Kamnitsas, K., Ferrante, E., Parisot, S., Ledig, C., Nori, A.V., Criminisi, A., Rueckert, D., Glocker, B., 2016. Deepmedic for brain tumor segmentation, in: International workshop on Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries, Springer. pp. 138–149.

Liu, J., Li, M., Wang, J., Wu, F., Liu, T., Pan, Y., 2014. A survey of mri-based brain tumor segmentation methods. Tsinghua Science and Technology 19, 578–595.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440.

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al., 2015. The multimodal brain tumor image segmentation benchmark (brats). IEEE Transactions on Medical Imaging 34, 1993–2024.

Menze, B.H., Van Leemput, K., Lashkari, D., Weber, M.A., Ayache, N., Golland, P., 2010. A generative model for brain tumor segmentation in multi-modal images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 151–159.

Milletari, F., Navab, N., Ahmadi, S.A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 Fourth International Conference on 3D Vision (3DV), IEEE. pp. 565–571.

Myronenko, A., 2018. 3d mri brain tumor segmentation using autoencoder regularization, in: International MICCAI Brainlesion Workshop, Springer. pp. 311–320.

Nyúl, L.G., Udupa, J.K., 1999. On standardizing the mr image intensity scale. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine 42, 1072–1081.

Ostrom, Q.T., Gittleman, H., Fulop, J., Liu, M., Blanda, R., Kromer, C., Wolinsky, Y., Kruchko, C., Barnholtz-Sloan, J.S., 2015. Cbtrus statistical report: primary brain and central nervous system tumors diagnosed in the united states in 2008-2012. Neuro-oncology 17, iv1–iv62.

Pereira, S., Pinto, A., Alves, V., Silva, C.A., 2016. Brain tumor segmentation using convolutional neural networks in mri images. IEEE Transactions on Medical Imaging 35, 1240–1251.

Pisapia, D.J., 2017. The updated world health organization glioma classification: cellular and molecular origins of adult infiltrating gliomas. Archives of pathology & laboratory medicine 141, 1633–1645.

Rohlfing, T., Zahr, N.M., Sullivan, E.V., Pfefferbaum, A., 2010. The sri24 multichannel atlas of normal adult human brain structure. Human brain mapping 31, 798–819.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

Tu, Z., Narr, K.L., Dollár, P., Dinov, I., Thompson, P.M., Toga, A.W., 2008. Brain anatomical structure segmentation by hybrid discriminative/generative models. IEEE Transactions on Medical Imaging 27, 495–508.

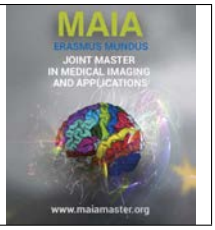
Wang, G., Li, W., Ourselin, S., Vercauteren, T., 2017. Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks, in: International MICCAI Brainlesion Workshop, Springer. pp. 178–190.

Wels, M., Carneiro, G., Aplas, A., Huber, M., Hornegger, J., Comaniciu, D., 2008. A discriminative model-constrained graph cuts approach to fully automated pediatric brain tumor segmentation in 3-d mri, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 67–75.

Yi, Z., Criminisi, A., Shotton, J., Blake, A., 2009. Discriminative, semantic segmentation of brain tissue in mr images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 558–565.

Medical Imaging and Applications

Master Thesis, June 2019



Automatic detection of calcification groups in DBT using domain adaptation from mammograms

Doiriel Vanegas C.[a], Mehmet Ufuk Dalmis[b], Michiel Kallenberg[b], Jaap Kroes[b]

[a] *University of Burgundy (France), UNICLAM (Italy), University of Girona (Spain)*
[b] *Screenpoint Medical (the Netherlands)*

Abstract

Breast cancer is one of the most common and dangerous cancers among women worldwide. An early detection leads to less invasive treatment options and increases the survival rate. For this reason, breast cancer screening programs have been established around the world, using mammography as the main imaging technique. One early sign of breast cancer in mammograms is the appearance of calcification groups. With the aim of increasing the sensitivity of the detection and reduce the workload created by screening programs, there is continuous interest in developing more accurate systems for automatic detection of the disease. Computer aided detection (CAD) algorithms have been developed for digital mammography (DM), and in recent years, the advent of deep learning has led to systems with a performance comparable to that of the average radiologist. On the other hand, digital breast tomosynthesis (DBT) is raising its popularity as screening technique, because it provides pseudo-3D images of the breast and reduces the masking effect of breast tissue on lesions. Due to the novelty of the technique, large databases of screening populations are not yet available to train deep learning algorithms. It would hold great value, if the information already collected from extensive mammography datasets could be used in the development of new algorithms for detecting breast cancer in DBT. Cross domain image to image translation has achieved state of the art performance in different problems for computer vision and could be helpful in this context. In this work, we investigated cycleGAN to generate DM-like images from DBT, with the purpose of using a classifier for microcalcifications previously trained on DM data. A dataset containing around 12,000 exams was used for training and validation. A cycleGAN was trained using unpaired images from Siemens DBT and Hologic DM. This resulted in the generation of realistic DM-like images providing an increase of 0.065 in the AUC of the classifier in DBT images. Additionally, we proposed to include the classifier inside the translation network during training, to guide it on making more reliable transformations. Although the loss of the classifier decreased during training, there was no significant improvement in the generated images. Further study should be focused on the integration of the new classifier network and the balance of the losses in the cycleGAN objective.

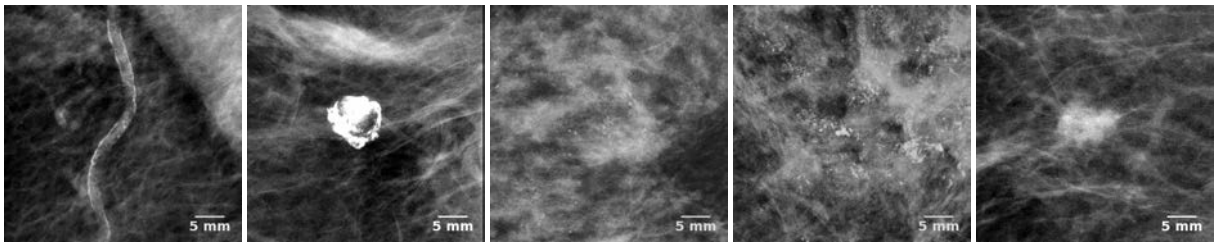
Keywords: Breast cancer, microcalcifications, digital breast tomosynthesis, deep learning, domain adaptation

1. Introduction

Statistics from the World Health Organization show breast is the most common cancer site and the deadliest cancer among women worldwide. According to a study conducted by Curado et al. (2007), one in eight women in the European Union will develop breast cancer before the age of 85. Fortunately, there is a good chance of recovery if the detection is made at an early stage.

Breast cancer is characterized by the uncontrolled growth of breast cells and the symptoms may include: a breast lump, changes in the size or appearance of the breast, discharge from the nipples or changes in the skin over the breast. The most common techniques to detect the disease are: breast examination, mammography, breast ultrasound, biopsy and breast magnetic resonance imaging (MRI).

Early detection of breast cancer is a key factor on prognosis, for this reason self-exam campaigns and national screening programs have been established. Breast



(a) Benign: vascular (b) Benign: popcorn (c) Benign: scattered (d) Malignant: scattered (e) Malignant: on mass

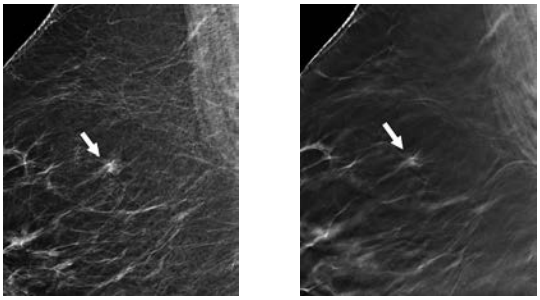
Figure 1: Examples of benign and malignant breast calcifications found in mammograms.

cancer screening in Europe has reduced the mortality rate by 25 - 31% for women invited for screening, and 38 - 48% for women actually screened (Broeders et al., 2012). However, this has meant an increase in the workload of radiology centers and radiologists.

Mammography has been for years the standard protocol for breast cancer screening, being able to detect about 80% to 90% of breast cancer cases in asymptomatic women. One important early sign of breast cancer in mammograms is the appearance of calcification groups (MCs), which are found in 30% to 50% of mammographically diagnosed cancer cases (Lanyi, 2012). Calcifications are calcium deposits that can be seen in mammograms as bright white spots, since calcium absorbs x-rays. Not all calcifications are signs of cancer, most macrocalfications (greater than 1mm) are caused by benign processes (Figures 1a, 1b), while microcalcifications (smaller than 1mm) require more attention, especially if they are clustered together as seen in Figures 1d, 1e. However, classification is not always trivial, as some benign calcifications can be small and clustered like in Figure 1c.

The sensitivity of mammography is impaired by the overlapping of tissues inherent to the technique, which makes more difficult the detection and diagnosis of abnormalities (Park et al., 2007). This effect is especially visible in women with dense breasts, since a higher portion of fibroglandular tissue is present. Breast density is also considered one of the risk factors for breast malignancy. MCs are difficult to see when they are overlapped with breast tissues, and superimposition of overlapping breast density can obscure them.

Digital breast tomosynthesis (DBT) has emerged as an evolution of digital mammography. DBT allows to obtain multiple projection views, acquired when an x-ray source rotates around the imaged breast. The projection images are subsequently reconstructed into several slices. Therefore, breast tissue volume can be visualized in sequential sections through the breast, reducing the masking effect of the fibroglandular tissue and improving breast cancer detection (Wu et al., 2003). Figure 2 shows an asymmetry as seen in both DM and DBT images. This asymmetry could be diagnosed as a malignant lesion by looking at the DM image only, but the DBT image reveals it is actually a benign lesion. Regarding MCs, DBT has shown a superior or equivalent visualization power to that of DM (Byun et al., 2017), (Andersson et al., 2008).

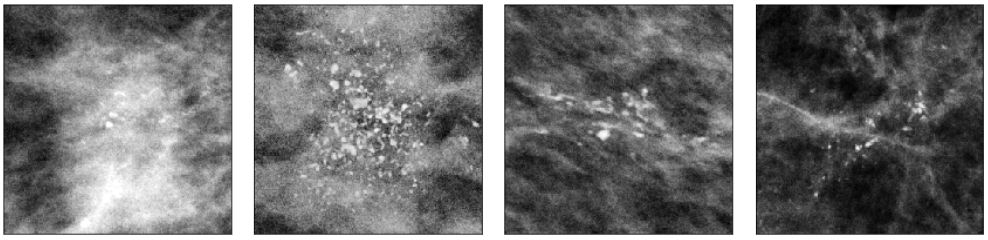


(a) DM (b) DBT

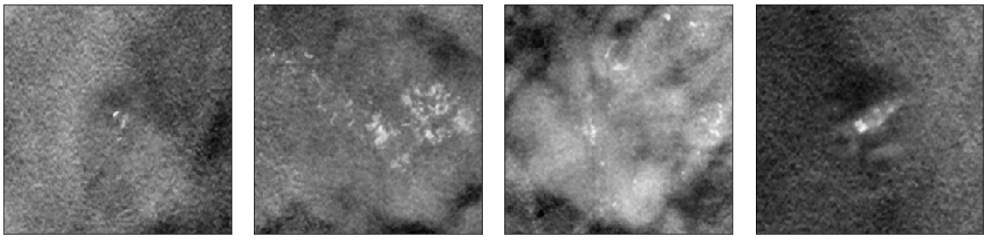
Figure 2: Digital mammography (DM) vs digital breast tomosynthesis (DBT): 2a close-up mediolateral oblique view in DM shows asymmetry in right upper breast (arrow); 2b DBT image reveals crossing Cooper ligaments and fibroglandular tissue (arrow) with no associated mass or spiculations. Recall was not necessary and this asymmetry was stable on routine follow-up screening (Hooley et al., 2017).

To improve the detection of breast cancer in screening and reduce the workload, a lot of research has been focused on the automatic detection of lesions in mammograms with computer aided detection (CAD) systems (Shen et al., 1993) (Strickland and Hahn, 1996), which in recent years have made new advances using machine learning and articial intelligence (Rodríguez-Ruiz et al., 2018) (Wu et al., 2019).

DBT is rising as the standard imaging technique for evaluation of breasts. Due to the increased number of image slices, the workload generated by this technique is significantly higher, and the need for accurate CAD systems becomes more urgent. Unfortunately, large public training DBT datasets for developing deep learning classifiers on screening populations are not yet available. Despite DBT is based on x-rays as DM, there are significant differences in image resolution, contrast and noise (Nelson et al., 2016). Figure 3 shows an example where DBT offers a better visualization of objects of certain size and contrast (medium and large microcalcifications), but provides poorer overall resolution



(a) DM images



(b) DBT images

Figure 3: Comparison between images containing microcalcifications generated by digital mammography (Hologic DM) and digital breast tomosynthesis (Siemens DBT). Differences in contrast and noise patterns can be observed.

and noise properties. Whereas the DM image contains white noise texture, the DBT image exhibits a mottled noise appearance. These variations prevent accurate algorithms trained on mammograms to succeed on DBT images.

It would be helpful if the information already available from extensive mammography datasets could be used somehow to train CAD systems for DBT. The success of domain adaptation on this setting, could open the possibility of using networks trained on mammograms directly on DBT images. In this work, we investigated the use of domain adaptation to improve the classification of calcification groups in DBT images, of a network previously trained on mammograms.

2. State of the art

2.1. Automated detection of microcalcifications

As in many areas of image analysis, deep learning techniques have achieved state of the art performance on the detection and classification of microcalcifications in mammograms. Wang et al. (2017) employed a CNN for detection of MCs in both screen-film and digital mammograms, outperforming their previously developed machine learning detector. Mordang et al. (2016) also trained a deep convolutional neural network (DCNN) on a much larger dataset, with images from three different vendors, achieving a higher sensitivity compared to cascade classifiers.

The detection of MCs in DBT poses a different challenge, since the clusters are often spread along several slices. A synthetic 2D projection image can be derived from the DBT study, which in combination to the DBT volume has a similar sensitivity and specificity to that of DM screening (Lai et al., 2018). 2D projection images are often used to simplify the problem of detection. Reiser et al. (2008) developed an algorithm for detection of microcalcifications using a linear classifier trained on projection images. Due to the limited amount of data, the parameters could not be optimized, and the sensitivity achieved was below the concurrent detection algorithms on DM data. Although few deep learning techniques can be found in the literature, Samala et al. (2016a) showed the potential of this approach when trained a DCNN using candidates extracted from the projection images obtaining a patch based AUC of 0.93.

It is clear that the amount of available data is a limitation for the development of deep learning based classifiers, since a large number of samples is required to adjust the high number of parameters in the model. To deal with this problem, several approaches can be considered: transfer learning, data augmentation, image to image translation, among others.

2.2. Transfer learning

The advantages of using transfer learning in medical imaging problems has been widely discussed in the literature, Tajbakhsh et al. (2016) demonstrated that the use of a pre-trained CNN with adequate fine-tuning outperformed or performed as well as a CNN trained from scratch, and fine-tuned CNNs were more robust to the size of training sets than CNNs trained from scratch. Shin et al. (2016) also showed it is possible to achieve state of the art performance by using transfer learning for thoraco-abdominal lymph node (LN) detection and interstitial lung disease (ILD) classification.

For the detection of masses in DBT, Samala et al. (2016b) proposed to train the generic layers of a DCNN using mammography and fine-tune the specific layers using DBT. Their results showed an increase in the AUC for classification of DBT images from 0.81 to 0.90. It would be interesting to see if similar results can be achieved without the need to create a separate classifier for DBT images, in other words, make DM classifiers work directly on DBT images.

2.3. Image synthesis for data augmentation

When considering data augmentation, the most common methods include rotation, scaling, translation, flipping and elastic deformation (Simard et al., 2003). These transformations often fail to capture the real variation in shape, size or location of specific pathologies, and certainly do not show the variations due to imaging protocols and/or sequences.

Generative adversarial networks (GANs) are deep neural network architectures able to estimate generative models via an adversarial process (Goodfellow et al., 2014). They consist of two networks trained simultaneously in a minimax two-player game: the first for image generation and the second to discriminate real from generated images. The model has achieved state of the art performance on many tasks in computer vision and has been used to generate synthetic medical data. Frid-Adar et al. (2018) employed deep convolutional GANs to generate synthetic patches of different liver lesions, and combined them to the real training data, obtaining an increase in both sensitivity and specificity. Costa et al. (2018) presented a GAN model able to synthesize new vessel networks and corresponding eye fundus images, using real pairs of images for training.

Korkinof et al. (2018) generated realistic high resolution mammography images using progressive GANs. The method consists in progressively increasing the resolution of generated images by gradually adding new layers to the generator and discriminator networks. However, the model was unable to generate calcification-like structures, and some artifacts were present in the generated images due to network failures.

According to the breast imaging report and database system (BI-RADS), the difference between benign and malignant microcalcifications is given by their distribution and morphology, therefore, the generation of images containing microcalcifications should be focused on reproducing these characteristics to generate realistic samples that can be used to increase the classification performance. The freedom given to general GAN approaches in the generation of new images can be problematic in the context of creating samples with microcalcifications, and a more restrictive method could produce better results.

2.4. Image to image translation

One of the most impressive applications of GANs are the image to image translation networks pix2pix (Isola et al., 2017) and cycleGAN (Zhu et al., 2017), for paired and unpaired data respectively. These networks provide cross domain synthesis, which is of great value in medical imaging since it can be used to reduce extra acquisition time of different modalities (e.g. convert MRI to CT images for bone delineation) or to generate new samples constrained by the anatomical structures present in the source image. Given that paired data is rare in clinical practice, unpaired image to image translation has raised interest in the medical field. Wolterink et al. (2017) showed that using unpaired data yields better results than using aligned data. They used cycleGAN to synthesize brain CT images from brain MR images, and the model was able to generate images that closely approximated reference CT images.

Although the research on the use of cycleGAN for medical images is still at an early stage, some promising results have been found. Chuquicusma et al. (2018) performed a visual Turing test showing that radiologists considered lung nodules generated by a GAN as real in 67% to 100% of the experiments. Chartsias et al. (2017) showed a 15% increase of the accuracy of segmentation when training with both real and synthesized cardiac MRI data from CT.

CycleGAN could be used for adapting classifiers previously trained on mammography images to correctly classify lesions on DBT images. In theory DBT images can be translated to DM images using cycleGAN and the translated images (DM-like) could be fed directly to the DM classifier. This concept is attractive as no retraining of the DM classifier would be needed, avoiding the risks of decreasing the performance on the real data. This is precisely the concept we explored on this master thesis: the use of cycleGAN to improve the performance of a mammography classifier on DBT data, for the detection of microcalcifications.

3. Material and methods

3.1. Baseline cycleGAN framework

To increase the performance of the classifier on DBT data, we used cycleGAN to translate DBT images to DM. CycleGAN consists of four networks: one generator for each image type (G_{DM} and G_{DBT}) and two discriminators (D_{DM} and D_{DBT}). Figure 4 illustrates the relationship between the components inside the framework. To translate from DBT to DM, three networks are involved: G_{DM} translates an image from DBT to DM, G_{DBT} translates the synthetic DM image back to its original domain (DBT), and D_{DM} tries to distinguish between synthesized and real DM images. While D_{DM} seeks to spot the fake mammography images, G_{DM} tries

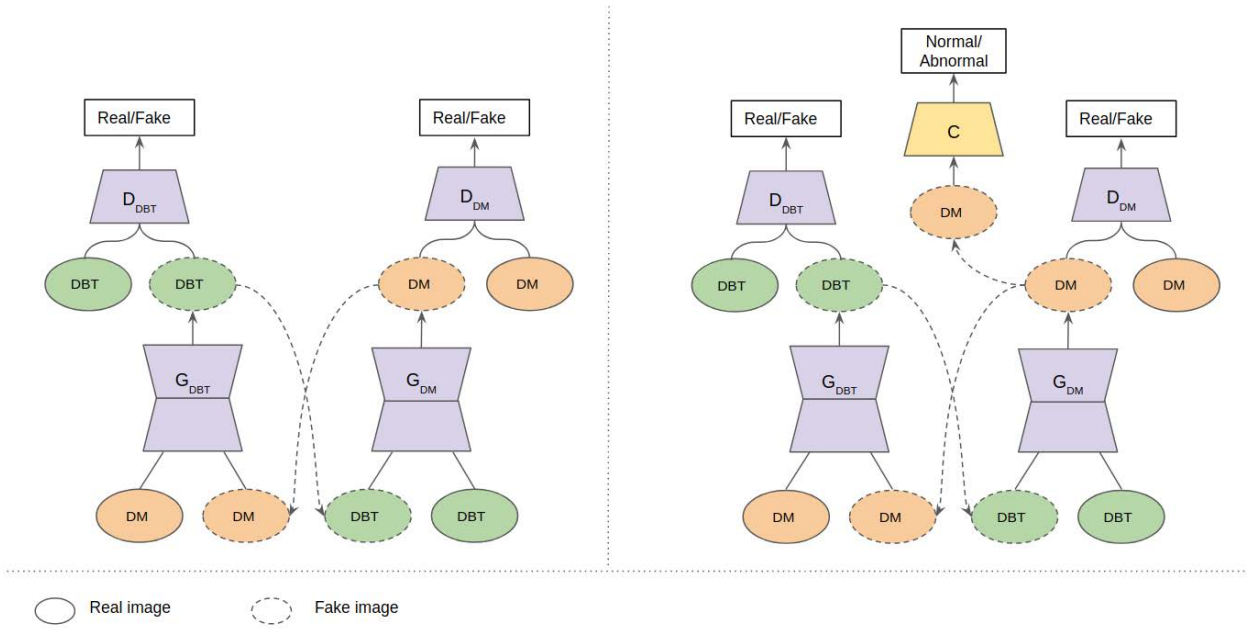


Figure 4: Frameworks used in this work to convert DBT images to DM images. Left: original cycleGAN. Right: proposed modification including a classifier trained on DM data (class-aware cycleGAN).

to prevent this by synthesizing more realistic images. For stability, the backward process is also trained: the same generators G_{DM} and G_{DBT} are used to go from DM to DBT and back, while the discriminator D_{DBT} tries to distinguish fake from real DBT images.

The adversarial goals of the discriminator and generator networks can be represented in the loss function: D_{DM} tries to predict the label 1 for real DM images while the label for the synthetic images should be 0. Therefore, in the generation of DM, the objective is to minimize Equation 1.

$$\mathcal{L}_{DM} = (1 - D_{DM}(I_{DM}))^2 + D_{DM}(G_{DM}(I_{DBT}))^2 \quad (1)$$

Similarly, to generate DBT from DM, Equation 2 is to be minimized.

$$\mathcal{L}_{DBT} = (1 - D_{DBT}(I_{DBT}))^2 + D_{DBT}(G_{DBT}(I_{DM}))^2 \quad (2)$$

The main difference between cycleGAN and the original GAN framework is the concept of *cycle consistency*. This is related to the property of the network to be able to return to the original domain. In human language translation, it can be expected that when translating a sentence from English to Spanish, and then translating the result back to English, a sentence close to the original is recovered. This is a way to ensure the translation is reliable and both generators are consistent.

Another reason for using cycle consistency in image to image translation, is the fact that regular adversarial training can learn mappings to generate random permutation of images in the target domain from a set of input images. In our case, such behaviour is not desirable, since the input and the output image must contain the same ground level information, e.g. same calcifications at the same locations, and only the style needs to be different. Adversarial losses alone cannot ensure that a given output image will resemble closely its source, and cycle consistency is used to reduce the space of possible mapping functions. The cycle consistency loss is defined by Equation 3.

$$\begin{aligned}\mathcal{L}_{cycle} = &\|G_{DBT}(G_{DM}(I_{DBT})) - I_{DBT}\|_1 + \\ &\|G_{DM}(G_{DBT}(I_{DM})) - I_{DM}\|_1\end{aligned} \quad (3)$$

The final cycleGAN objective is given by the summation of the previously defined losses as described in Equation 4. Where λ controls the relative importance of the objectives.

$$\mathcal{L} = \mathcal{L}_{DM} + \mathcal{L}_{DBT} + \lambda\mathcal{L}_{cycle} \quad (4)$$

3.2. Class-aware cycleGAN

The main problem in this project is the classification of abnormal microcalcification patterns from normal ones. To ensure the translated images remain classifiable, we proposed to include the original class of the image to guide the learning process. This was implemented by changing the final objective to include the loss of a microcalcifications classifier resulting in Equation 5. A schematic of this framework can be found in Figure 4.

$$\mathcal{L} = \mathcal{L}_{DM} + \mathcal{L}_{DBT} + \lambda_1\mathcal{L}_{cycle} + \lambda_2\mathcal{L}_{classifier} \quad (5)$$

3.3. Classifier previously trained on mammography data

The classification between normal and abnormal samples was assessed using a CNN previously trained for classification of microcalcifications in DM data. The network follows a VGG-11 architecture shown in Figure 5; batch normalization, padded convolutions and ReLU activations were used. The input to this network is a gray scale image patch, with a bit-depth of 16, and dimensions 224×224 pixels, which corresponds to approximately 2×2 cm.

Figure 5: Architecture of the classifier for microcalcifications in DM patches.

3.4. Dataset

A large set of unpaired DBT and DM images were used for the experiments, containing normal and abnormal exams. The images came from two vendors: Siemens for DBT images, and Hologic for DM images (Table 1). Exams were collected at multiple clinical centers across Europe, including sites in the Netherlands, Germany, and the UK. Data collection sites are representative for regular breast cancer screening and asymptomatic patients in hospitals who have mammograms for a variety of reasons; such as increased risk for breast cancer or not being invited for population-based screening program (e.g. because of age under 50). For the inclusion of the normal exams in the dataset a follow-up of at least one year was required. Most of the exams have mediolateral oblique (MLO) and craniocaudal (CC) views of both the left and right breast. The ground truth was based on the clinical reports from both radiology and pathology. A region was considered a true positive if the center of the patch falls within the annotated contour of the lesion.

Table 1: Number of clinical exams available.

	Normal exams	**Abnormal exams**
Hologic DM	180,406	1,378
Siemens DBT	674	60

In the experiments, the translation was done from Siemens DBT images to Hologic DM, because the majority of the images for training the classifier came from the Hologic vendor.

3.5. Patch preparation

Patches containing candidates of calcification groups were extracted from an synthetic 2D projection image. In order to construct the synthetic image, relevant points in the DBT volume were identified by a classifier that aimed at detecting individual calcifications; subsequently these points were mapped to a 2D representation. After extraction the patches were preprocessed using the method from Kooi et al. (2017), and then center cropped to a size of 224 x 224 pixels to be used as input to the cycleGAN network. Note that the dimensions of the patches were selected to match the input of the classification network.

Table 2 refers to the patches used during the experiments. Note that in Table 1, a significantly higher amount of normal exams for Hologic DM were available, therefore, more than 240,000 patches were extracted from them. To reduce the imbalance between patches of different domains and decrease computational time, 10,000 patches were randomly selected and used for the experiments. The patches were split into training (80%) and validation (20%) sets.

Table 2: Number of patches used for training and validation.

	Normal	**Abnormal**
Hologic DM	10,000	2,920
Siemens DBT	654	102

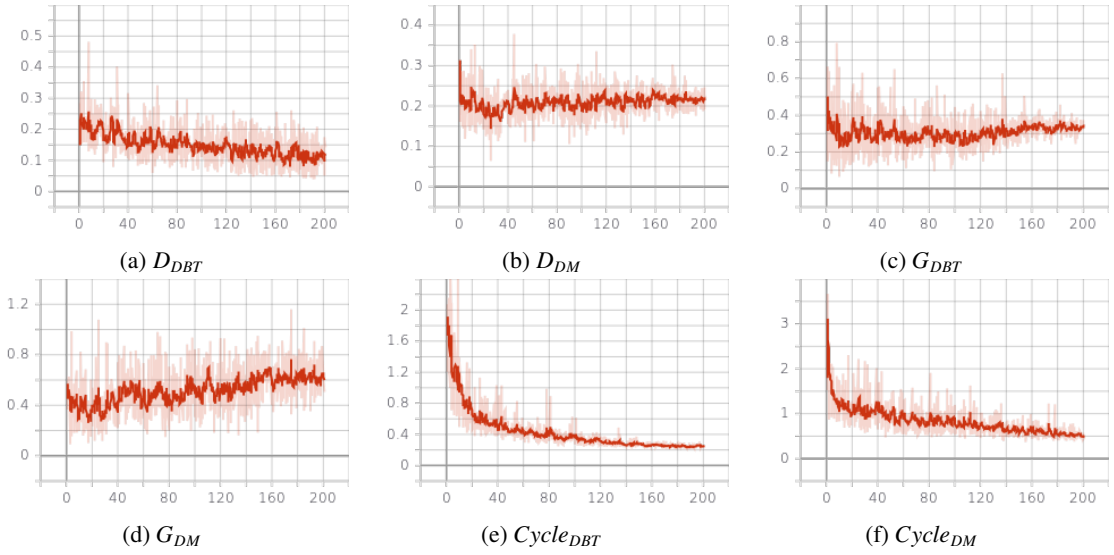
For further evaluation, an independent test set was available, the distribution of extracted patches is described in Table 3.

Table 3: Number of patches in the test set.

	Normal	**Abnormal**
Hologic DM	2387	125
Siemens DBT	1207	151

3.6. Augmentations and upsampling

For training the cycleGAN, the following data augmentations were used: random vertical flip, random 90° rotations, and small random translations. As the dataset included fewer abnormal samples, for the experiments when normal and abnormal patches were used, interleaving was employed to upsample the abnormal patches and avoid class imbalance.



(a) D_{DBT} (b) D_{DM} (c) G_{DBT}

(d) G_{DM} (e) $Cycle_{DBT}$ (f) $Cycle_{DM}$

Figure 6: Training losses for baseline cycleGAN.

3.7. Evaluation

As the goal of this project was to improve the classification of MCs, the main evaluation method used was the analysis of the receiver operating characteristic (ROC) curve. More specifically the area under the curve (AUC) was employed to select the best models. For the experiments with the baseline cycleGAN framework, the translation models were saved every five epochs of training, then, the models were used to translate the DBT images in the validation set to DM, and finally, the AUC of the classifier on the translated images was measured. The model providing the highest AUC was selected.

For the class-aware cycleGAN framework, the evaluation was done with cross validation. In contrast to the baseline framework, where only normal samples were used for training and validation could be done in all the abnormal samples; the class-aware framework required both normal and abnormal samples for training, reducing the samples left for validation. To have an estimate of the performance for all the abnormal samples, the training was performed in five folds and the scores were pooled. The data in Table 2 was split into five, for each fold the resulting model was saved every five epochs and the best model was chosen. Once all models were available, the scores of the validation data for each fold were concatenated to calculate the AUC on the totality of the validation set. For the testing set all models were tested and the average AUC was calculated.

4. Results

4.1. Baseline cycleGAN framework

The original Pytorch implementation of cycleGAN from Zhu et al. (2017), was adapted to work with 16 bit depth, gray scale images. CycleGAN was trained using normal DBT patches, and during testing time, both normal and abnormal patches were transformed into mammography-like patches, which were then fed to the classifier of microcalcifications. The training was done for 200 epochs with a batch size of 4, and learning rate 0.0002. The parameter in Equation 4 was set to $\lambda = 10$. After 100 epochs the learning rate was linearly decreased to zero.

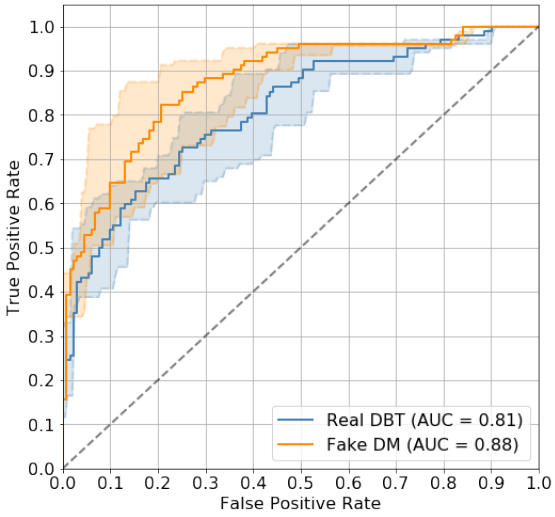
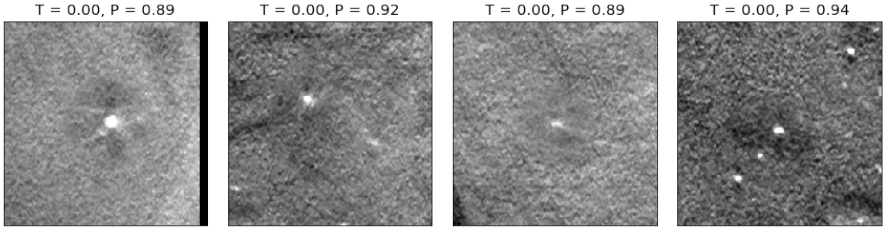
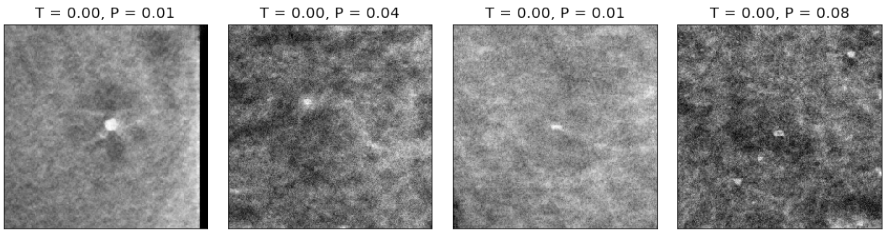


Figure 7: **Baseline cycleGAN framework**. Receiver operating characteristic curves of the microcalcifications' classifier for different inputs in the validation set. Real DBT: real tomosynthesis patches. Fake DM: mammography-like patches generated via original cycleGAN.

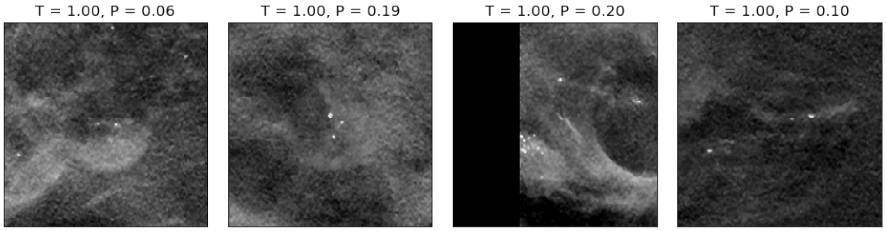
The change of the different losses during training can be seen in Figure 6, as the curves were very noisy a smoothing (bold line) was applied to better visualize the trend over time. This experiment took 79 hours to complete on a single NVIDIA Titan V GPU. The translation



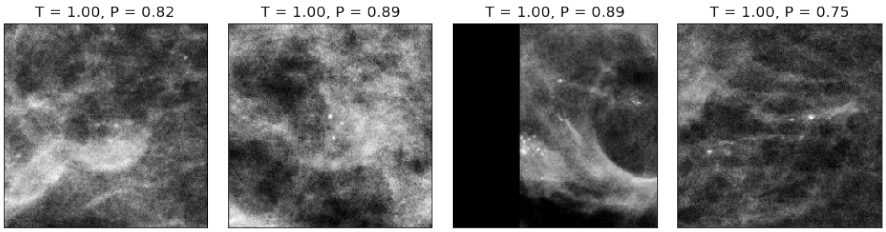
(a) Normal patches: original DBT



(b) Normal patches: DM-like generated



(c) Abnormal patches: original DBT



(d) Abnormal patches: DM-like generated

Figure 8: **Baseline cycleGAN framework**. Examples of correctly translated DBT patches to DM. T: true label, P: predicted label.

model which provided the highest AUC was chosen to draw Figure 7. In the validation set, the model achieved an AUC of 0.88 (95% confidence interval: 0.82 to 0.91, estimated from 1000 bootstrap samples), with a mean AUC increase of 0.065 (95% confidence interval: 0.001 to 0.132) with respect to the performance of the real DBT patches.

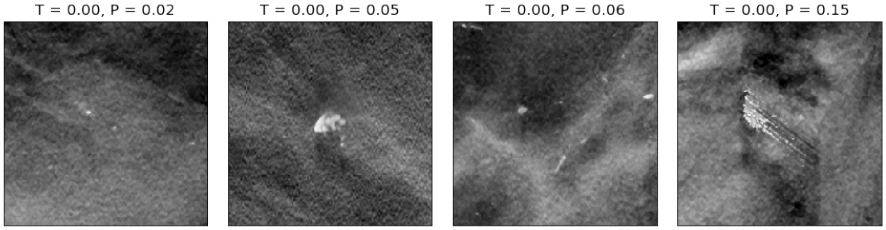
The model was evaluated on the test set, giving the results shown in Figure 13. The performance of the classifier on real mammography data is also displayed for comparison. The AUC of the classifier on the translated patches via the baseline framework was 0.06 higher than on real DBT.

Figure 8 shows examples of DBT images from the test set correctly translated to the DM domain. For each image, the classification score before and after translation is shown. The success of the framework is illustrated in the reduction of the difference between the true and predicted scores for the translated images, and in the noticeable changes in contrast and noise patterns.

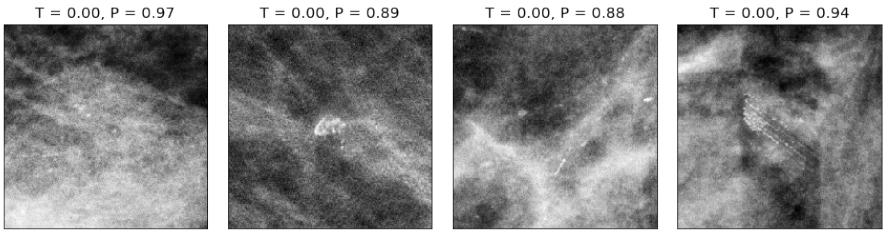
Figure 9 shows examples where the translation decreased the ability of the classifier to correctly label the images. This evidences how the score given by the classifier is affected by small changes in the image appearance, such as a decreased contrast between calcifications and the surrounding tissues.

4.2. Class-aware cycleGAN framework

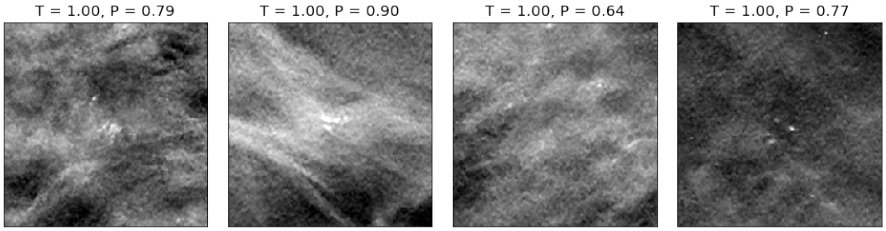
The best model obtained from the baseline experiments was fine-tuned using the class-aware cycleGAN framework. The training was done in five folds for 15 epochs, with a batch size of 4, and learning rate 0.0002. The parameters in Equation 5 were set to $\lambda_1 = 10$ and



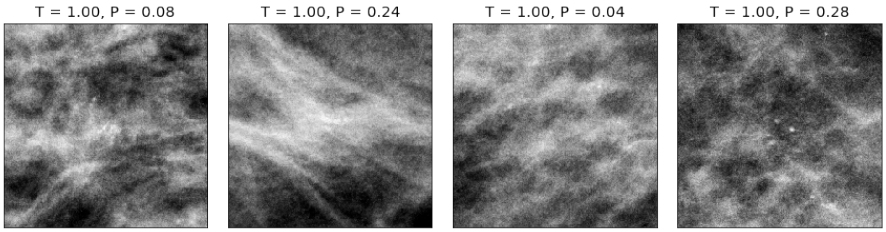
(a) Normal patches: original DBT



(b) Normal patches: DM-like generated



(c) Abnormal patches: original DBT



(d) Abnormal patches: DM-like generated

Figure 9: **Baseline cycleGAN framework**. Misclassified patches using translated images. T: true label, P: predicted label.

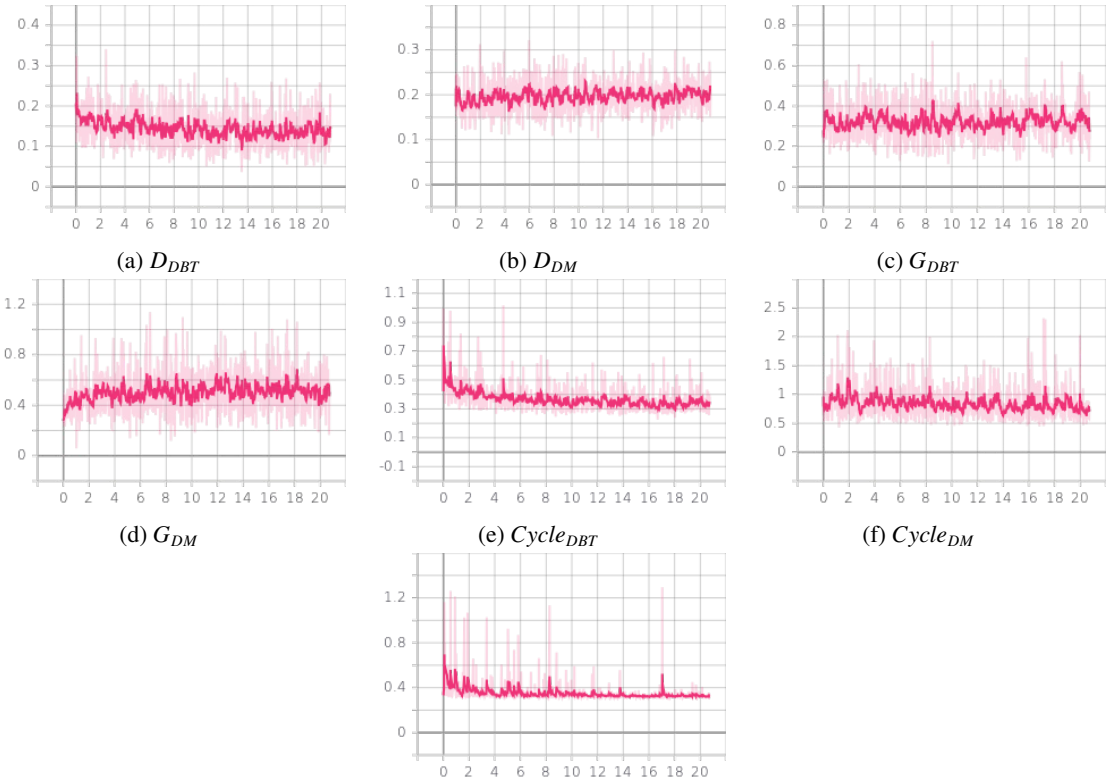
$\lambda_2 = 1$. Model training took 21 hours per fold on a single NVIDIA Titan V GPU.

Figure 10 shows an example of the losses during training of one fold, which is representative of the behaviour of the other four folds.

The images in the validation set translated using the class-aware cycleGAN had an AUC of 0.74 (95% confidence interval: 0.688 to 0.787, estimated from 1000 bootstrap samples), with a mean AUC decrease of 0.058 with respect to the performance of the real DBT patches (95% confidence interval: -0.124 to 0.004), as shown in Figure 11.

The ROC curves of the testing set images translated via the different models obtained in different folds are shown in Figure 12. A variation in performance between models is observed, with an AUC between 0.71 and 0.79. The average performance of the class-aware cycleGAN is shown in Figure 13, along with the performance of original DM patches, original DBT patches and fake DM patches generated via the baseline cycleGAN framework. The class-aware cycleGAN did not represent an improvement in AUC with respect to the baseline framework, having a similar performance to that of the images without translation.

Examples of images that benefited from the class-aware cycleGAN are shown in Figure 14, while unsuccessful translations can be seen in Figure 15. For both cases the contrast and noise properties translated via cycleGAN are visible, but in general the images tend to look smoother and even blurred. The translation of a sample image via the different models obtained from cross validation is given in Figure 16, showing the small changes that each model made in the final image.



(a) D_{DBT} (b) D_{DM} (c) G_{DBT}

(d) G_{DM} (e) $Cycle_{DBT}$ (f) $Cycle_{DM}$

(g) $Classifier$

Figure 10: Training losses for the class-aware cycleGAN framework.

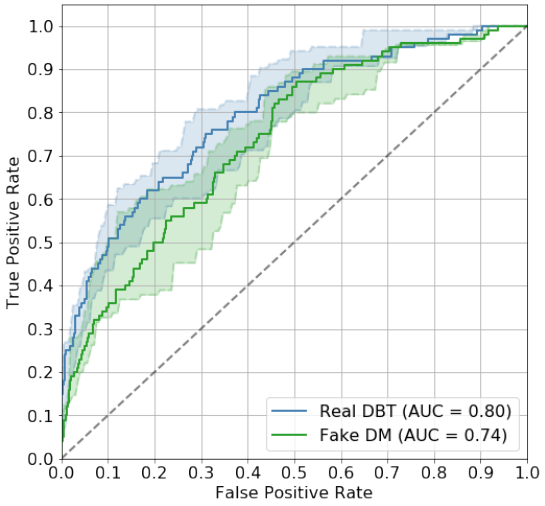


Figure 11: **Class-aware cycleGAN**. Receiver operating characteristic curves of the microcalcifications' classifier on the validation set for different inputs. Real DBT: real tomosynthesis patches. Fake DM: mammography-like patches generated via class-aware cycleGAN.

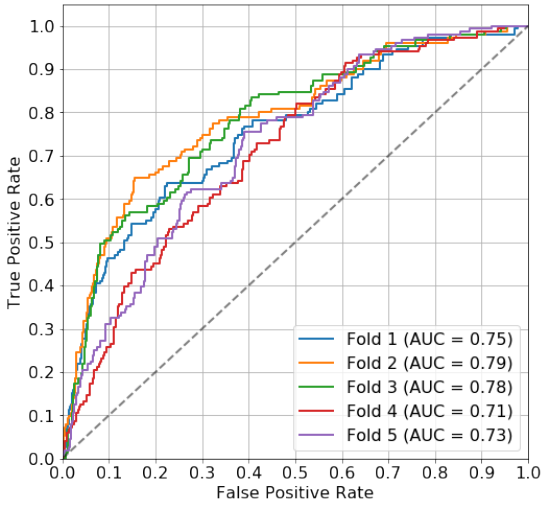


Figure 12: **Class-aware cycleGAN**. Receiver operating characteristic curves of the microcalcifications' classifier on the test set for different folds.

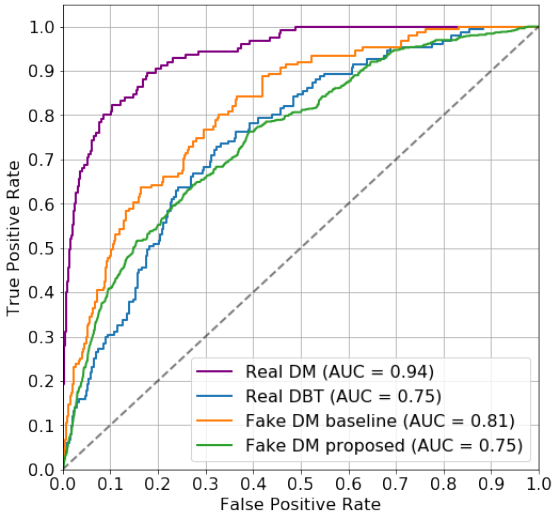


Figure 13: Receiver operating characteristic curves of the microcalcifications' classifier on the test set for different inputs. Real DM: real mammography patches. Real DBT: real tomosynthesis patches. Fake DM baseline: mammography-like patches generated via original cycleGAN. Fake DM proposed: mammography-like patches generated via the class-aware cycleGAN .

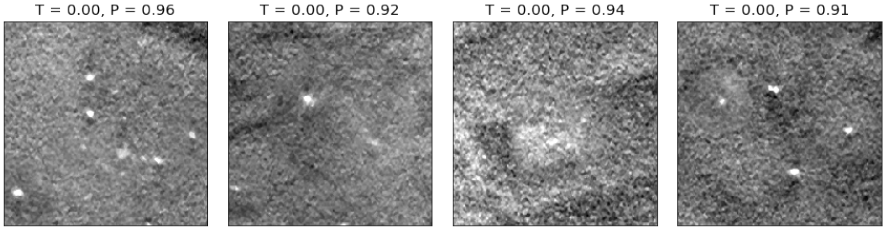
5. Discussion

In this work we evaluated the use of cycleGAN for transferring the style from mammography to DBT images, aiming to improve the classification of microcalcifications on DBT images. As described in Section 2, the use of image to image translation for medical images is of great interest for the field. However, the methods employed to assess the performance of these approaches are often qualitative, measuring if an expert can believe that the generated images are realistic. Such evaluation is relevant in computer vision where the appearance of the images is the most important, but in medical images the image intensity of each pixel has semantic meanings, such as the x-ray attenuation. For this reason it is not only important that the image looks realistic but also that its shapes and structures are preserved. CycleGAN often gives certain freedom to the network to generate the images and this can lead, in some cases, to non-optimal results for diagnosis or classifier training. Since we wanted to improve the classification of microcalcifications, in addition to an image appearance perspective, we evaluated the performance of the classifier in the fake images.

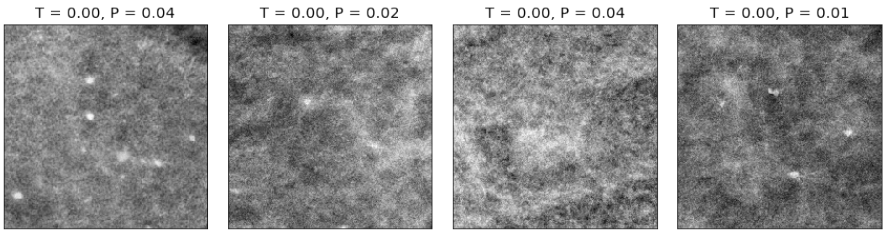
The available dataset was unpaired and imbalanced. DBT and DM exams may come from different patients without any relation to each other, this is the reason why cycleGAN was chosen over pix2pix. An important issue for training the image to image translation network was to decide what ratio of data to use. As the main focus of this work is on microcalcifications, it was important that the lesions were preserved and translated accurately. A first approach was to perform the training using only abnormal samples (see Appendix A), but the limited amount of data was an obstacle for improvement. Using both normal and abnormal samples for training, without any supervision in the original cycleGAN framework, can lead to normal patches trying to be translated to abnormal ones and vice versa. Fortunately, normal patches came from a candidate selection algorithm, and they also contained calcifications (see Figure 8a), in principle this information could be enough to train the translation network. Training with normal patches also increases the size of the dataset for DBT images significantly, which improved the performance.

The training losses in Figure 6 showed the expected behaviour: all losses changed gradually as the training advanced. A training is considered to fail when the losses decrease rapidly and reach zero, making no changes on the images. The discriminator for DM images (Figure 6b) decreased during the first few epochs, it then stabilized and tended to increase slightly during the rest of the training, showing the generator was creating more realistic images and fooling the discriminator. The generator's loss (Figure 6d) showed a tendency to increase, meaning the training was close to completion and the generation could not be improved further. On the other hand, the behaviour of the translation from DM to DBT seemed to be poorer; the discriminator of DBT images (Figure 6a) kept decreasing and the generator's loss (Figure 6c) tended to remain constant. This can be explained with the imbalance of the dataset: more examples of DM images were available for the network to learn how should be their appearance. The cycle consistency losses (Figures 6f, 6e) decreased significantly with each iteration and this behaviour was maintained until the training stopped. This can be an indication that the weight assigned to the cycle losses was higher than optimal, and the network was more interested in generating images that could be translated back, than in accurate translations. This suggests a decrease in the weight of the cycle consistency loss could potentially improve the image generation.

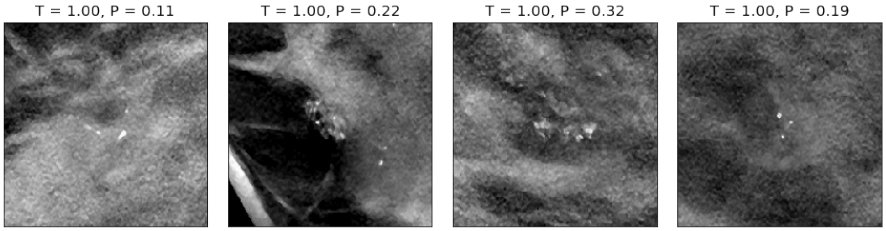
The results revealed that the original cycleGAN framework was able to generate mammography like images from DBT. As shown in Figures 8 and 9, the contrast and noise patterns of the DM images were successfully transferred to the DBT images, providing realistic results. For some images in Figure 9 it can be seen that although the overall contrast of the image was improved, the visibility of calcifications was compromised, resulting in a harder classification. The ROC analysis in Figure 7 shows an overall improvement in classification when the fake images were fed to a classifier previously trained on mammography data. In Figure 13 the comparison between the use of real and translated images as input to the classifier is displayed. However,



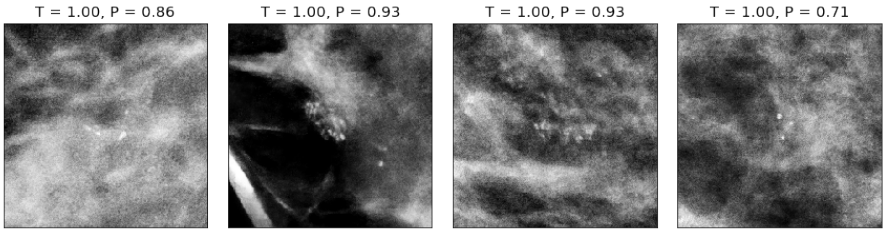
(a) Normal patches: original DBT



(b) Normal patches: DM-like generated



(c) Abnormal patches: original DBT



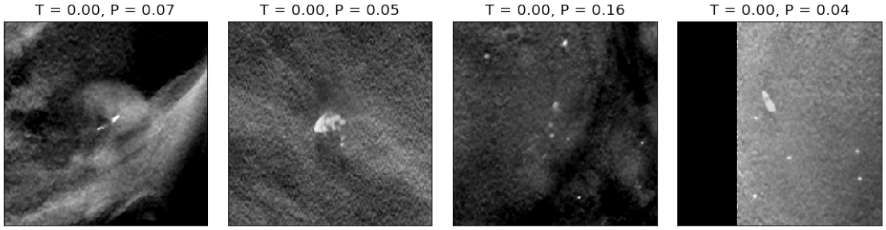
(d) Abnormal patches: DM-like generated

Figure 14: **Class-aware cycleGAN**. Examples of correctly translated DBT patches to DM. T: true label, P: predicted label.

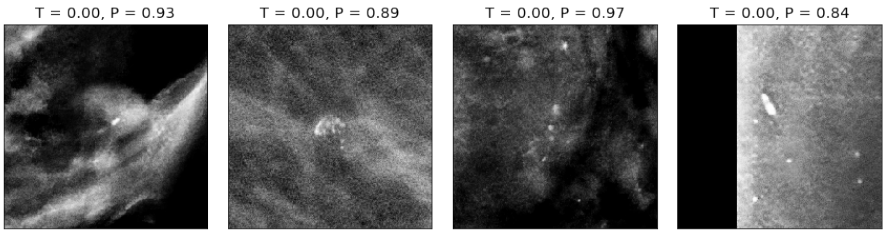
this comparison is not completely fair, since DM and DBT images were not paired. In Appendix B, an additional classifier was trained on combined DBT and DM data, demonstrating that using DM-like images translated via the baseline cycleGAN framework on a classifier trained on DM data, produced a similar AUC to the one obtained using original DBT images on a classifier trained with both DM and DBT images (Figure B.20).

Using the translated images on the classifier was expected to provide a result close to the original DM data, this notion was supported by the appearance of the generated images. However, as seen in Figure 13, there is still some room for improvement. The original cycleGAN framework does not apply much restriction to the generation of the images, and in clinical data too much freedom can be detrimental. The main constraint used in cycleGAN is the cycle consistency, which forces the original image to be reconstructed from the fake one. This has been proven problematic since the network learns to cheat by encoding the information of the original image in the generated version in a way that is not easily visible (Chu et al., 2017). In our case, we saw that when few data were available for training, as the training advanced, the reconstructed images (from the fake) became more similar to the original, but the quality of the translation was not necessarily improved. Examples of this can be found in Figure A.19, the first row shows how some calcifications were faded, and the second, a change in the pattern of the calcifications. Interestingly, for both cases the reconstructed images look almost identical to the original ones.

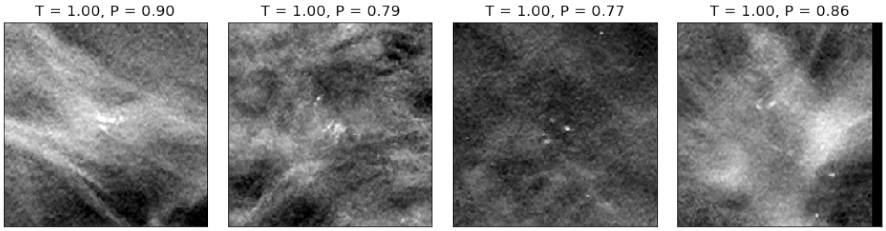
This problem of cycleGAN has been discussed in the literature and some alternatives have been proposed to limit the freedom of the network during image gen-



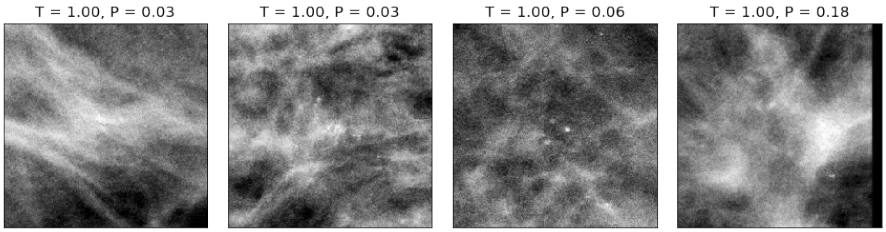
(a) Normal patches: original DBT



(b) Normal patches: DM-like generated



(c) Abnormal patches: original DBT



(d) Abnormal patches: DM-like generated

Figure 15: **Class-aware cycleGAN**. Misclassified patches using translated images via the class-aware cycleGAN. T: true label, P: predicted label.

eration, for example conditional cycleGAN (Lu et al., 2018). This solution is however not desirable for our problem since the generation requires an attribute, such as the class label, and this is precisely what is unknown at testing time. We proposed to use the classifier inside the cycleGAN to guide the translation, and generate images that are not only realistic but also classifiable. This concept was implemented via the loss function, by adding the classifier's loss to the cycleGAN objective.

The training losses of the class-aware cycleGAN in Figure 10, display a relatively stable behaviour. The classifier's loss tended to decrease, specially during the first epochs of fine-tuning, showing this was when the network benefited the most from the addition of the new loss. When its value decreases, it may be helpful to increase the weight assigned to the classifier's loss as the training advances. This would help to provide more meaningful information to the cycleGAN and improve the image translation. Cross validation results show a wide variation in performance for the models obtained in the different folds (Figure 12). Figure 16 shows how variations of the models can affect brightness and contrast of microcalcifications in the translated images. This can be a product of the limited number of abnormal training samples, despite the efforts to upsample them during training. When combining the scores of all validation images from every fold, the result was a decrease in performance (see Figure 11). Due to the variation between the models generated on each fold (see ROC curves in Figure 12 and images in Figure 16), the distribution of scores given by the classifier to the images translated using each model can differ. When the scores were pooled together, these variations were not taken into account to search for an appropriate classification

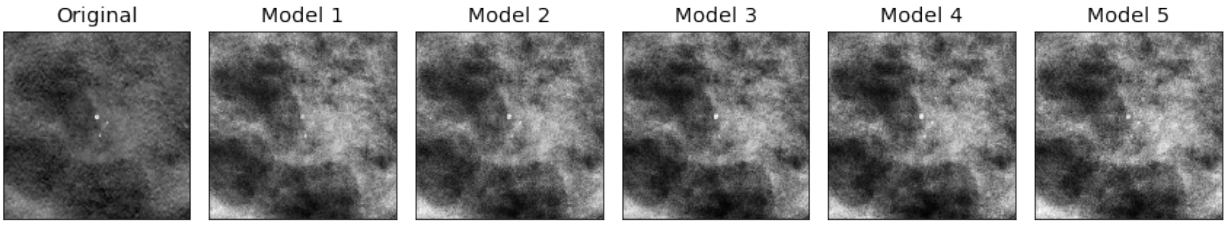


Figure 16: **Class-aware cycleGAN**. A sample DBT patch translated to DM using the models obtained from different splits in cross validation.

threshold, producing a lower AUC. The decrease in performance was confirmed in the test set (see Figure 13), showing that the use of the classifier to guide the translation reduced the positive effects of image generation via cycleGAN. A closer look to the generated images in Figure 15, shows an over smoothing effect in the translated images that can be responsible of the poor performance.

As with other GANs, stabilizing the training of cycleGAN is a complex task, and adding a new loss to the objective increases the complexity of the model. The optimal weight of this new loss must be further investigated, since it can have an impact in the final image. A small weight can make the network ignore the information from the classifier and a very high loss can result in limiting the generation networks too much in producing the images.

In the first stage of this project, we used a microcalcifications' classifier, trained on heavily augmented data in order to have a high performance on both DM and DBT images. Using DBT images translated to DM on this classifier was not showing an improvement regardless of the efforts to fine-tune hyperparameters. It seemed that cycleGAN had little to offer in this setting, so we decided to simplify the experiments using a less crafted classifier to evaluate the use of cycleGAN, leading to the results described above. It is possible that the better classifier was already well trained, and translating DBT images to a DM-like space decreased its performance. Probably, since the fake images were neither completely DBT nor DM, the classifier had problems to understand them. In contrast, the classifier used afterwards felt more comfortable with images that resembled more of a DM than images from a different domain.

6. Conclusions

In this work we investigated the application of cycleGAN to improve the classification of microcalcifications on DBT images. DBT patches of suspicious regions were translated to the DM domain, aiming to improve their scores on a classifier previously trained on DM data. The results suggest this methodology is promising as the AUC for classification on DBT images increased in 0.065, being almost as good as a classifier trained on DBT data. To further improve these results, we proposed to provide more guidance to the cycleGAN framework by including the classifier's loss during training. However, it was problematic to assess the weight of this new loss in regard to the multiple loss functions inside the cycleGAN framework, leading to an overall decrease of performance. Further work should be focused on tuning the weight for the different losses in the cycleGAN objective, specially for the cycle consistency and the classifier.

7. Acknowledgments

I would like to thank Screenpoint Medical for providing all the materials, infrastructure and support for the development of this thesis, specially to my supervisors for all their loving guidance.

References

Andersson, I., Ikeda, D.M., Zackrisson, S., Ruschin, M., Svahn, T., Timberg, P., Tingberg, A., 2008. Breast tomosynthesis and digital mammography: a comparison of breast cancer visibility and birads classification in a population of cancers with subtle mammographic findings. European radiology 18, 2817–2825.

Broeders, M., Moss, S., Nyström, L., Njor, S., Jonsson, H., Paap, E., Massat, N., Duffy, S., Lynge, E., Paci, E., 2012. The impact of mammographic screening on breast cancer mortality in europe: a review of observational studies. Journal of medical screening 19, 14–25.

Byun, J., Lee, J.E., Cha, E.S., Chung, J., Kim, J.H., 2017. Visualization of breast microcalcifications on digital breast tomosynthesis and 2-dimensional digital mammography using specimens. Breast cancer: basic and clinical research 11, 1178223417703388.

Chartsias, A., Joyce, T., Dharmakumar, R., Tsaftaris, S.A., 2017. Adversarial image synthesis for unpaired multi-modal cardiac data, in: International Workshop on Simulation and Synthesis in Medical Imaging, Springer. pp. 3–13.

Chu, C., Zhmoginov, A., Sandler, M., 2017. Cyclegan, a master of steganography. arXiv preprint arXiv:1712.02950 .

Chuquicusma, M.J., Hussein, S., Burt, J., Bagci, U., 2018. How to fool radiologists with generative adversarial networks? a visual turing test for lung cancer diagnosis, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE. pp. 240–244.

Costa, P., Galdran, A., Meyer, M.I., Niemeijer, M., Abràmoff, M., Mendonça, A.M., Campilho, A., 2018. End-to-end adversarial retinal image synthesis. IEEE transactions on medical imaging 37, 781–791.

Curado, M.P., Edwards, B., Shin, H.R., Storm, H., Ferlay, J., Heanue, M., Boyle, P., 2007. Cancer incidence in five continents, Volume IX. IARC Press, International Agency for Research on Cancer, Lyon, France.

Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., Greenspan, H., 2018. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. Neurocomputing 321, 321–331.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Advances in neural information processing systems, pp. 2672–2680.

Hooley, R.J., Durand, M.A., Philpotts, L.E., 2017. Advances in digital breast tomosynthesis. American Journal of Roentgenology 208, 256–266.

Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 5967–5976.

Kooi, T., Litjens, G., Van Ginneken, B., Gubern-Mérida, A., Sánchez, C.I., Mann, R., den Heeten, A., Karssemeijer, N., 2017. Large scale deep learning for computer aided detection of mammographic lesions. Medical image analysis 35, 303–312.

Korkinof, D., Rijken, T., O'Neill, M., Yearsley, J., Harvey, H., Glocker, B., 2018. High-resolution mammogram synthesis using progressive generative adversarial networks. arXiv preprint arXiv:1807.03401 .

Lai, Y.C., Ray, K.M., Lee, A.Y., Hayward, J.H., Freimanis, R.I., Lobach, I.V., Joe, B.N., 2018. Microcalcifications detected at screening mammography: synthetic mammography and digital breast tomosynthesis versus digital mammography. Radiology 289, 630–638.

Lanyi, M., 2012. Diagnosis and differential diagnosis of breast calcifications. Springer Science & Business Media.

Lu, Y., Tai, Y.W., Tang, C.K., 2018. Attribute-guided face generation using conditional cyclegan, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 282–297.

Mordang, J.J., Janssen, T., Bria, A., Kooi, T., Gubern-Mérida, A., Karssemeijer, N., 2016. Automatic microcalcification detection in multi-vendor mammography using convolutional neural networks, in: International Workshop on Breast Imaging, Springer. pp. 35–42.

Nelson, J.S., Wells, J.R., Baker, J.A., Samei, E., 2016. How does c-view image quality compare with conventional 2d ffdm? Medical physics 43, 2538–2547.

Park, J.M., Franken Jr, E.A., Garg, M., Fajardo, L.L., Niklason, L.T., 2007. Breast tomosynthesis: present considerations and future applications. Radiographics 27, S231–S240.

Reiser, I., Nishikawa, R., Edwards, A., Kopans, D., Schmidt, R., Papaioannou, J., Moore, R., 2008. Automated detection of microcalcification clusters for digital breast tomosynthesis using projection data only: a preliminary study. Medical physics 35, 1486–1493.

Rodríguez-Ruiz, A., Krupinski, E., Mordang, J.J., Schilling, K., Heywang-Köbrunner, S.H., Sechopoulos, I., Mann, R.M., 2018. Detection of breast cancer with mammography: effect of an artificial intelligence support system. Radiology 290, 305–314.

Samala, R.K., Chan, H.P., Hadjiiski, L.M., Cha, K., Helvie, M.A., 2016a. Deep-learning convolution neural network for computer-aided detection of microcalcifications in digital breast tomosynthesis, in: Medical Imaging 2016: Computer-Aided Diagnosis, International Society for Optics and Photonics. p. 97850Y.

Samala, R.K., Chan, H.P., Hadjiiski, L.M., Helvie, M.A., Wei, J., Cha, K.H., 2016b. Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography. Medical physics 43 12, 6654.

Shen, L., Rangayyan, R.M., Desautels, J.L., 1993. Detection and classification of mammographic calcifications. International Journal of Pattern Recognition and Artificial Intelligence 7, 1403–1416.

Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M., 2016. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging 35, 1285–1298.

Simard, P.Y., Steinkraus, D., Platt, J.C., 2003. Best practices for convolutional neural networks applied to visual document analysis, in: Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings., pp. 958–963. doi:10.1109/ICDAR.2003.1227801.

Strickland, R.N., Hahn, H.I., 1996. Wavelet transforms for detecting microcalcifications in mammograms. IEEE Transactions on Medical Imaging 15, 218–229.

Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2016. Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE transactions on medical imaging 35, 1299–1312.

Wang, J., Nishikawa, R.M., Yang, Y., 2017. Global detection approach for clustered microcalcifications in mammograms using a deep learning network. Journal of Medical Imaging 4, 024501.

Wolterink, J.M., Dinkla, A.M., Savenije, M.H., Seevinck, P.R., van den Berg, C.A., Išgum, I., 2017. Deep mr to ct synthesis using unpaired data, in: International Workshop on Simulation and Synthesis in Medical Imaging, Springer. pp. 14–23.

Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., Jastrzebski, S., Févry, T., Katsnelson, J., Kim, E., et al., 2019. Deep neural networks improve radiologists' performance in breast cancer screening. arXiv preprint arXiv:1903.08297 .

Wu, T., Stewart, A., Stanton, M., McCauley, T., Phillips, W., Kopans, D.B., Moore, R.H., Eberhard, J.W., Opsahl-Ong, B., Niklason, L., Williams, M.B., 2003. Tomographic mammography using a limited number of low-dose cone-beam projection images. Medical Physics 30, 365–380.

Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. 2017 IEEE International Conference on Computer Vision (ICCV) , 2242–2251.

Appendix A. Training with abnormal data only

As a first approach to the problem discussed in this work, we considered to train the baseline cycleGAN using only abnormal images. The experiment was run for 100 epochs with a batch size of 1, and learning rate 0.0002. The parameter in Equation 4 was set to $\lambda_1 = 10$. The model was saved every five epochs and the AUC was calculated on the validation set. The model which provided the highest AUC was chosen to draw Figure A.17. The model achieved an AUC of 0.77 (95% confidence interval: 0.681 to 0.843, estimated from 1000 bootstrap samples), with a mean AUC increase of 0.012 with respect to the performance of the real DBT patches (95% confidence interval: -0.081 to 0.118). The comparison of the performance of this approach against training with only normals is shown in Figure A.18 for the test set.

Appendix B. Classifier trained on combined DBT and DM data

An additional classifier for MCs with the architecture described in Section 3.3 was trained using combined DBT and DM patches. Figure B.20 shows the comparison of the results obtained using the different classifiers and types of images from the testing set. The AUC of the DBT images translated to DM via cycleGAN was comparable to that of the original DBT images classified using a network trained on DBT images.

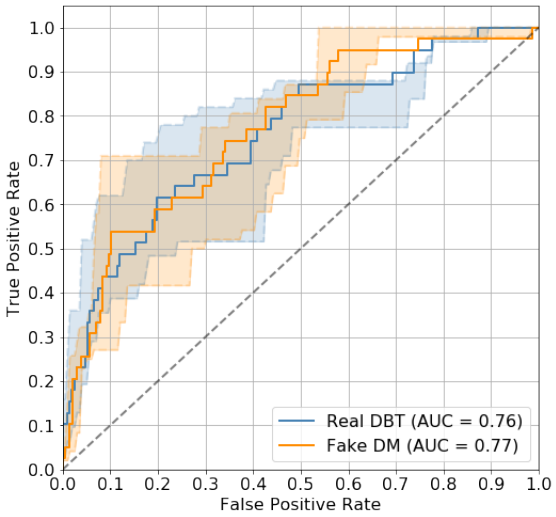
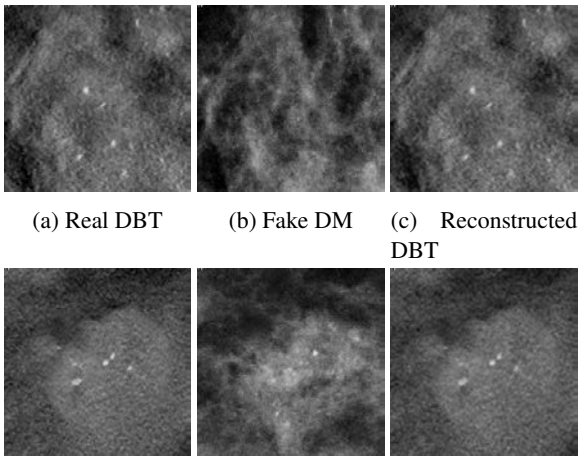


Figure A.17: Receiver operating characteristic curves of the micro-calcifications' classifier for different inputs. Real DBT: real tomosynthesis patches. Fake DM: mammography-like patches generated via original cycleGAN, using a model trained on abnormal patches.



(a) Real DBT (b) Fake DM (c) Reconstructed DBT

(d) Real DBT (e) Fake DM (f) Reconstructed DBT

Figure A.19: Example of calcifications poorly translated but recovered during reconstruction.

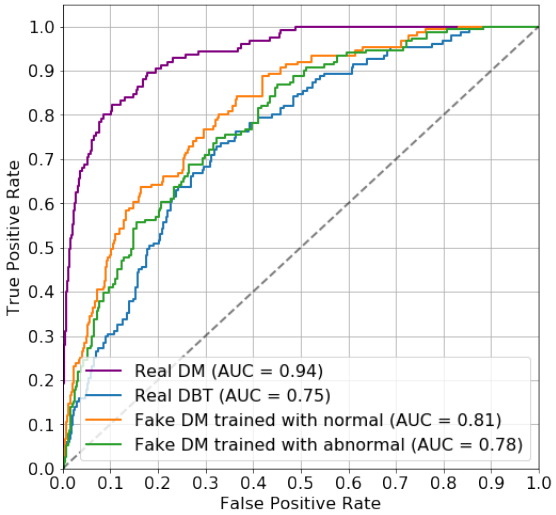


Figure A.18: Receiver operating characteristic curves of the micro-calcifications' classifier for different inputs. Real DM: real mammography patches. Real DBT: real tomosynthesis patches. Fake DM: mammography-like patches generated via original cycleGAN.

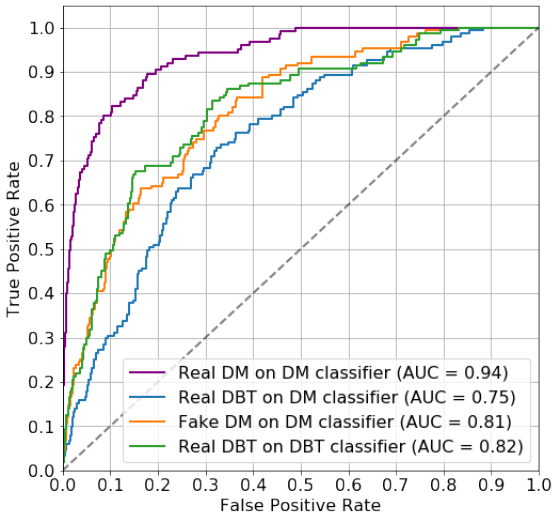
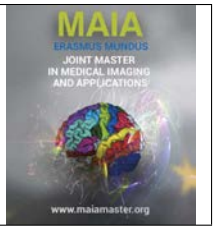


Figure B.20: Receiver operating characteristic curves comparing calcification classifiers with different inputs. Real DM on DM classifier: real DM patches classified with a network trained on DM. Real DBT on DM classifier: real DBT patches classified with a network trained on DM. Fake DM on DM classifier: DM-like patches generated via original cycleGAN classified with a network trained on DM. Real DBT on DBT classifier: real DBT patches classified with a network trained on DBT and DM.

Medical Imaging and Applications

Master Thesis, June 2019



Automated Background Parenchymal Enhancement Classification in Breast DCE-MRI

Ama Katseena Yawson[a,b,c], Oliver Diaz[c], Robert Martí[c]

[a] *Université de Bourgogne(France)*
[b] *Università degli studi di Cassino e del Lazio Meridionale(Italy)*
[c] *Universitat de Girona(Spain)*

Abstract

Breast cancer is the most common neoplastic disease in women around menopause. Due to this alarming risks, detection at an early stage is key in reducing the rate of mortality. Background parenchymal enhancement (BPE) is the enhancement of fibroglandular tissue (FGT) of the breast in response to MRI contrast agent. Current studies proves that BPE can be used as a biomarker to determine the risk of developing breast cancer. However, BPE rating suffers from large intra- and inter-observer variability. The purpose of this study to investigate the use of automated tools (traditional machine learning and deep learning) to classify BPE into their respective classes (mild, minimal, moderate and marked). The study was conducted using 491 patients' study. Each of these cases were manually evaluated by 3 radiologists from 3 different countries. The qualitative approach was assessed as an initial step to establish ground-truth labels for the automated techniques. Prior to BPE classification, segmentation of the region of interest (i.e FGT) was carried out using an in-house proprietary segmentation tool. In the traditional machine learning approach, hand-crafted features were extracted from the FGT in both pre-contrast (t0) and post-contrast (t1) DCE-MRI volumes. These set of features were automatically classified into the 4 BPE classes using SVM and Random Forest (RF) classifiers. Unlike the traditional machine learning approach, inputs to the deep neural network were 2D slices selected from the middle of each volume for both t0 and t1. Using the concept of transfer learning, pre-trained Resnet-50 model from PyTorch archive was used to automatically extract and classify features. The optimal classifier found in the traditional machine learning technique was the RF classifier with 100 trees. Combination of all extracted features showed an overall accuracy and F1_score of 50% and 0.46 respectively. Comparatively, the results obtained for the deep learning technique was higher than the RF classifier with an overall accuracy and F1_score of 58% and 0.55 respectively. Thus, machine learning algorithms have the potential to help automate BPE classification and provide supplementary opinion to radiologists. However, more evaluation is needed before introducing it in a clinical environment.

Keywords: Breast cancer, DCE-MRI, FGT, BPE, Biomarker, Traditional Machine learning, Deep learning

1. Introduction

Breast cancer is the most common neoplastic disease in women around menopause (Kamińska et al., 2015). According to estimates from the American Cancer Society (ACS) for breast cancer in the United States, approximately 41,760 breast cancer death are expected to occur among women in 2019 (Siegel et al., 2019). Though the exact cause of this type of cancer is unknown, it can be attributed mainly to factors such as age, family history, amount of dense tissue, hormonal changes, lifestyle, etc. As a result of this alarming risks, detection at an early stage is key in reducing the rate of mortality. The commonly used imaging modal-

Email address: ykatseena@yahoo.com (Ama Katseena Yawson)

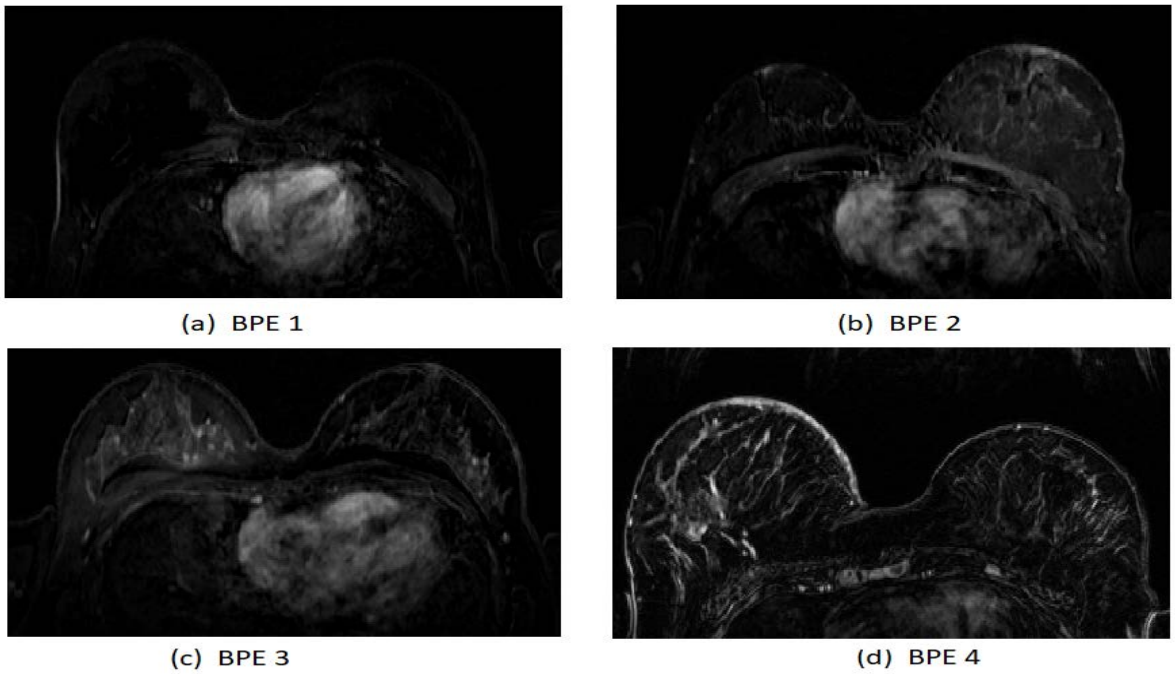


Figure 1: Axial DCE-MRI slices of the first post-contrast difference images (t1-t0) of 4 different women showing the 4 categories of BPE: (a) BPE 1 - minimal enhancement, (b) BPE 2 - mild enhancement, (c) BPE 3 - moderate enhancement, (d) BPE 4 - marked enhancement.

ity for early breast cancer detection is x-ray mammography. The use of other modalities such as magnetic resonance imaging (MRI) and tomosynthesis are rapidly evolving. Currently, ultrasound and MRI for breast cancer is recommended by the ACS as an adjunct to mammography for screening (Wu et al., 2016). In the conventional breast MRI protocols, dynamic contrast enhanced MRI (DCE-MRI) represents the most sensitive breast imaging technique for cancer detection. DCE-MRI consists of 3D acquisitions of the entire breast volume at different time points. Typically, a pre-contrast (t0) and post-contrast (t1,t2,...tn) volumes before and after the administration of an intravenous contrast agent. The time-signal intensity curves of the different post-contrast sequences reflect the dynamic signal intensity variations induced by uptake of contrast agent over a period of time and can be described by contrast enhancement kinetics (Kuhl et al., 1999)(Wu et al., 2016).

Background parenchymal enhancement (BPE) is the enhancement of fibroglandular tissue (FGT) of the breast in response to MRI contrast agent (Wu et al., 2016). Presently, BPE is evaluated both manually and qualitatively by radiologists using 4 ordinal categories defined in the breast imaging-reporting and data system (BI-RADS) as: minimal, mild, moderate and marked (Park et al., 2007)(Morris, 2007). Figure 1 illustrates sample DCE-MRI slices of the 4 BPE categories. While mammographic breast density has been established as an independent risk factor, current studies proves that BPE can be used as a biomarker to determine the risk of developing breast cancer, although they must be interpreted with caution (Dontchos et al., 2015)(Felix et al., 2016)(Mema et al., 2018). Likewise, it is associated with tumor characteristics, diagnostic performance and therapy response. However, evidence suggests that BPE correlates negatively with patients age and increases with greater hormonal activity (Müller-Schimpfle et al., 1997)(Pfleiderer et al., 2004)(King et al., 2011). In contrast, recent studies contradictorily suggest BPE to be an imaging feature without increased cancer coincidence in asymptomatic or high-risk patients (Baltzer et al., 2011)(Bennani-Baiti et al., 2016)(You et al., 2018).

King et al. (2011) were one of the first authors who examined the relationship between BPE and breast cancer risk. They presented evidence that BPE might be more sensitive to breast cancer odds than the amount of glandular tissue for certain women subsets. Nevertheless, further research is currently needed to explore the role played by BPE in cancer risk assessment and provide an unbiased mechanism for BPE quantification.

The most frequent BPE categories fall in the range of minimal or mild with slow early and persistent delayed kinetic curve (Giess et al., 2014). These characteristics make MRI interpretation di-

rect and easy. However, cases classified as moderate or marked may interfere in accurate differentiation of small breast lesions leading to increased false-positive rates and reduced sensitivity of MR examinations (DeMartini et al., 2012)(Giess et al., 2014). This effect may also lead to unnecessary biopsies and can influence a woman's choice on getting mastectomy rather than breast conservation therapy (Klifa et al., 2011).

Although the 4 BPE categories are clearly described by the American College of Radiology (ACR), BPE rating suffers from large intra- and inter-observer variability which ranges widely from fair to substantial (Observers' agreement, κ = 0.36–0.70) (Dontchos et al., 2015)(Pujara et al., 2018). This variability can be linked to various factors such as level of experience, acquired training and many others. Additionally, this task is tedious and time consuming. To reduce these limitations, automatic techniques are needed to aid radiologists in their final decisions. Such tools could have other potential applications, including training of young radiologists in BPE classification or even the reduction of inter- and intra-reader variability.

2. State of the art

In recent clinical scenarios, dedicated workstations are commonly used to aid radiologists in making their final decision for detection and recognition of breast lesions in DCE-MRI examinations (Gubern-Mérida et al., 2015). As the potential role of BPE in breast cancer risk determination has gained attention, attempts to quantify BPE has now become popular (Pujara et al., 2018). This popularity can be attributed to the inherent subjective, large intra- and inter-observer variability in the current state of evaluation by radiologist (Dontchos et al., 2015)(Pujara et al., 2018). In order to address these limitations, there is the need to develop automated medical analysis tools and computer-aided detection systems to aid in the interpretation of DCE-MRI breast examinations (Gubern-Mérida et al., 2015). These automated techniques have the ability to reduce the workload of radiologists and help to improve diagnosis.

A new technique for computing FGT enhancement in breast DCE-MRI was presented by Klifa et al. (2011) aimed at quantifying the enhancement of BPE. The adopted quantitative approach for measuring breast MRI enhancement was analyzed for a population of 16 healthy volunteers. Their algorithm was tested on high risked women who have already undergone 3 months of tamoxifen therapy. Quantitative parenchymal enhancement measures were made in all cases. From their experiment, they observed that high risk patient demonstrated a 37% mean reduction in background enhancement with treatment and hence suggested that quantitative methods are robust and promising tool that may allow radiologists to correctly quantify and document the potential adverse effect of BPE on diagnostic accuracy in larger populations.

Yang et al. (2015) published a new quantitative image analysis method for improving breast cancer diagnosis using DCE-MRI examinations. The aim of their work was to examine the feasibility of applying a novel quantitative method to aid in improving breast cancer diagnosis performance using DCE-MRI and integrating BPE features into the decision making process. Using a computer aided detection system, segmentation was made on the region of interest (ROI) i.e. FGT. From the ROI, 18 kinetic features were computed. 6 of these features were selected from the segmented breast tumour and the rest from the parenchymal regions (excluding the tumor). A support vector machine (SVM) classifier was trained and optimized using different combinations of the extracted features. The leave-one-case-out validation method was used to test the performance of the classifier and also assessed using area under the curve (AUC) of the Receiver Operating Characteristic (ROC). They concluded that quantitative BPE features provide useful knowledge to the kinetic features of breast tumours in DCE-MRI and hence their integration to computer-aided diagnosis techniques could improve breast cancer diagnosis based on DCE-MRI examinations.

Quantitative three dimensional assessment of FGT and BPE using a semi-automated computerized methods was developed by (Ha et al., 2016). In their approach, three-dimensional BPE quantification method was evaluated with standardized BPE qualitative cases. Based on their observation, there was a significant positive correspondence between quantitative MRI FGT assessment and qualitative MRI FGT and hence the quantitative technique may become a valuable tool in clinical use by providing computer generated standardized measurements with less intra- and inter-observer variability.

Similarily, Pujara et al. (2018) also made a comparison between qualitative and quantitative assessment of BPE on breast DCE-MRI. Qualitative evaluation was made by 4 breast radiologists using the 4-point scale (a–d same as BPE 1 - BPE 4) after the administration of contrast agent at the first 90 s and 180 s. With the aid of a phantom-validated segmentation algorithm, FGT masks were generated and co-registered to pre- and post-contrast fat suppressed images to delineate the region of interest and obtain a quantitative BPE measure (Pujara et al., 2018). ROC analyses and kappa coefficients (κ) were used as comparison metric be-

tween the subjective and quantitative approach. Using ROC analyses, the authors concluded that BPE at 90s was best predicted by the quantitative BPE approach compared to subjective assessment. However, at higher levels of quantitative BPE, agreement between subjective BPE and quantitative BPE significantly decreased for all four radiologists at 90 s and for 3 out of 4 radiologists at 180 s.

Despite the fact that quantitative methods proposed by Klifa et al. (2011), Yang et al. (2015), Ha et al. (2016) and Pujara et al. (2018) provide supplementary information to aid radiologist measure accurately FGT and BPE, their methods are still time-consuming and require initial delineation of the ROI by the radiologist which may introduce potential subjective bias (Eyal et al., 2009)(Clendenen et al., 2013)(Ha et al., 2019). Considering the recent breakthrough of deep learning algorithms in object detection and objection recognition, this has influenced its application in many medical fields. Ha et al. (2019) recently developed a fully automated approach using the convolutional neural network (CNN) for quantification of FGT and BPE. Using t0, t1, and (t1-t0) difference images of 137 patients, they manually segmented FGT and classified BPE levels to generate ground truth labels. With the aid of a new 3D CNN built from the standard 2D U-Net architecture, voxel-wise prediction for the whole breast and FGT borders was developed and implemented. At the end of their study, they successfully quantified FGT and BPE within an average of 0.42 s per MRI case. However, their approach is still limited in terms of validation.

In this study, machine learning techniques (traditional machine learning and deep learning) are investigated to classify BPE into their respective classes; aiming to alleviate the subjective nature of BPE. The study will be conducted using 405 patients' study. Each of these cases were manually evaluated by 3 radiologists from 3 different countries. The qualitative approach will be assessed as an initial step to establish ground-truth for the automatic methods. Evaluation will be provided in terms of overall accuracy, accuracy per class, F1_score, kappa (κ) agreement and the significance level (p-value). The whole pipeline will be carried out using Matlab®R2018b with Weka-3-8 (Traditional Machine Learning) and PyTorch 1.0.1.post2 (Deep learning).

3. Material and methods

3.1. Data Acquisition

Breast DCE-MRI volumes used in this study were collected from an existing clinical database at the Radboud University Medical Centre (Nijmegen, the Netherlands). The dataset consists of 491 DCE-MRI examinations acquired from 405 women who underwent breast cancer diagnosis at the hospital. In contrast to other studies, unilateral mastectomy and breast implant cases with large visibility of parenchymal enhancement were included in the dataset. However, bilateral mastectomy were excluded. DCE-MRI examinations were acquired using 3 different Siemens scanners(Magnetom Vision, Magnetom Avanto and Magnetom Trio) with magnetic field of 1.5 or 3 Tesla. Each of these scanners has a dedicated breast coil which includes control panel (CP) Breast Array, Siemens and Erlangen respectively. Flip angles of 20°, repetition time of 5.5 s and echo time of 1.7 s were employed. In each examination, TWIST (Time-resolved angiography With Stochastic Trajectories) sequence was followed as depicted in Figure 2. At least 3 post-contrast acquisitions were available in the dataset used in this study. The different scanners produced different set of MRI volume sizes of $256 \times 128 \times 112$, $384 \times 192 \times 160$, $448 \times 448 \times 160$, $448 \times 448 \times 176$ and $512 \times 256 \times 120$ with pixel spacing between 0.7 and $1.3mm^2$ and coronal slice thickness between 1 and 1.5 mm.

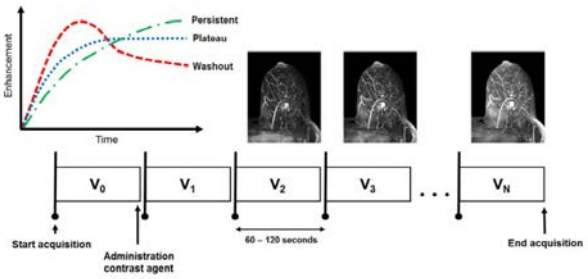


Figure 2: TWIST sequence for Breast DCE-MRI. 3D volume of the breast is taken before and after the administration of intravenous contrast agent. Signal enhancement computed over the ROI can be classified as persistent (green), plateau (blue) and early washout (red) (Gubern-Mérida et al., 2015).

3.2. Radiologist Assessment

As a way of addressing the bias that comes along with an individual reader, three expert radiologists from three different countries manually annotated BPE levels of the dataset. The detailed information of each reader is summarised in Table 1. Using the guidelines defined in the ACR BI−RADS MRI lexicon classification categories, each reader rated the level of BPE independently into the 4 ordinal categories. For this task, each reader visualised the maximum intensity projection images at time point t0 and t1 only to rate the BPE level. The pre-contrast and post-contrast time points were used

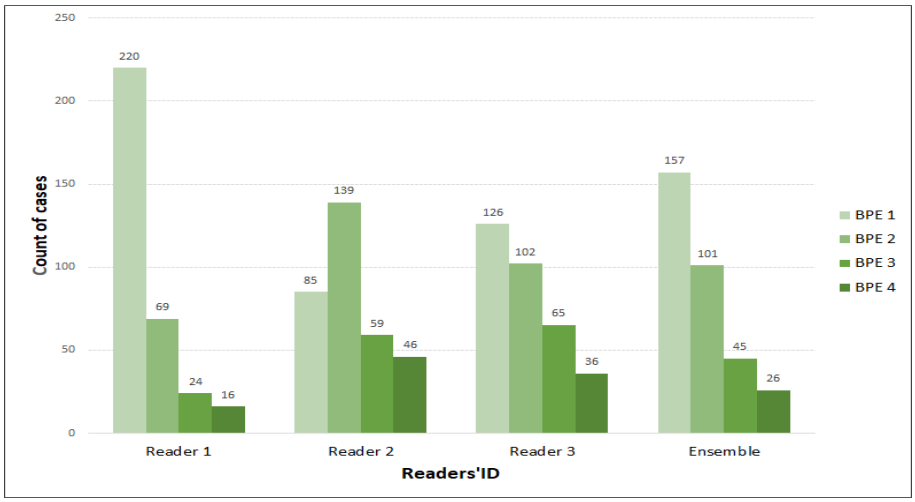


Figure 3: . Histogram of BPE annotations for each of the readers. The 4 plot represents the rate of each BPE reader with the last representing the fused rate for all the readers as described in the text.

because they are assumed to be representative for the BPE level within the entire volume (King et al., 2011); although it is still not clear which time point yields useful information for breast cancer risk stratification and prediction of response to treatment (Melsaether et al., 2017). As correctly predicted by Giess et al. (2014), all readers annotated most of the dataset as mild or minimal with few cases classified as moderate or marked. The rate of each reader is summarized in the histogram plot shown in Figure 3.

Table 1: Readers' details

Readers' ID	Country	Experience	Speciality
Reader 1 (R1)	Netherlands	8	Breast
Reader 2 (R2)	Germany	3	Breast
Reader 3 (R3)	Spain	25	Prostate

The agreement between readers was measured using the quadratic weighted cohen's kappa coefficient (κ) given by Equation 1.

$$\kappa = 1 - \frac{\Sigma_{i=1}^{n}\Sigma_{j=1}^{n} w_{ij}x_{ij}}{\Sigma_{i=1}^{n}\Sigma_{j=1}^{n} w_{ij}m_{ij}}, \quad (1)$$

where n is the number of classes and w_{ij}, x_{ij}, m_{ij} are elements in the weight, observed, and expected matrices respectively.

$\kappa < 0.0$ was interpreted as poor agreement, $0.0 \leq \kappa \leq 0.20$ as slight agreement, $0.20 < \kappa \leq 0.40$ as fair agreement, $0.40 < \kappa \leq 0.60$ as moderate agreement and $\kappa > 0.60$ as substantial agreement (Landis and Koch, 1977). Before computing the inter-observer agreement, cases with suboptimal segmentation mask were excluded. Similarily, datasets without any agreement with at least 2 readers were also filtered out. That is, patient study where the 3 readers had 3 independent rate (R1 - BPE 1, R2 - BPE 4 and R3 - BPE 2). This precaution was taken to avoid wrong assignment of labels; thus the dataset reduced from 491 to 329. Using the majority voting ensembling technique, the rate of the 3 readers were fused together to establish groundtruth labels for the automated approach. The ensemble rate for all the readers is also displayed in Figure 3. Afterwards, the agreement between the ensembled rate and the individual rates of each reader was also investigated.

3.3. FGT Segmentation

As an initial step to the machine learning approaches, segmentation of the FGT was carried out using an in-house proprietary segmentation tool developed by (Gubern-Merida et al., 2014). In the segmentation algorithm, atlas based segmentation was first used to delineate the borders of breast from the image background and the chest wall. The subsequent FGT segmentation involved expectation-maximization (EM) for Gaussian mixture models. EM was used to determine a threshold intensity value to distinguish FGT from adipose tissue. N4 bias-field correction algorithm was performed prior to segmentation. The final FGT segmentation was obtained within the breast volume after applying the fuzzy-c-means (FCM) segmentation algorithm. Segmentation was performed only in (t0) volumes and used for all the time points. Segmentation results were generally satisfactory but in some cases they were manually corrected to minimise labelling errors. The output of the segmentation was then used as a mask to extract appropriate voxel's information in non-bias-field corrected (t0) and (t1) volumes. Some samples of segmented FGT are displayed in Figure 4.

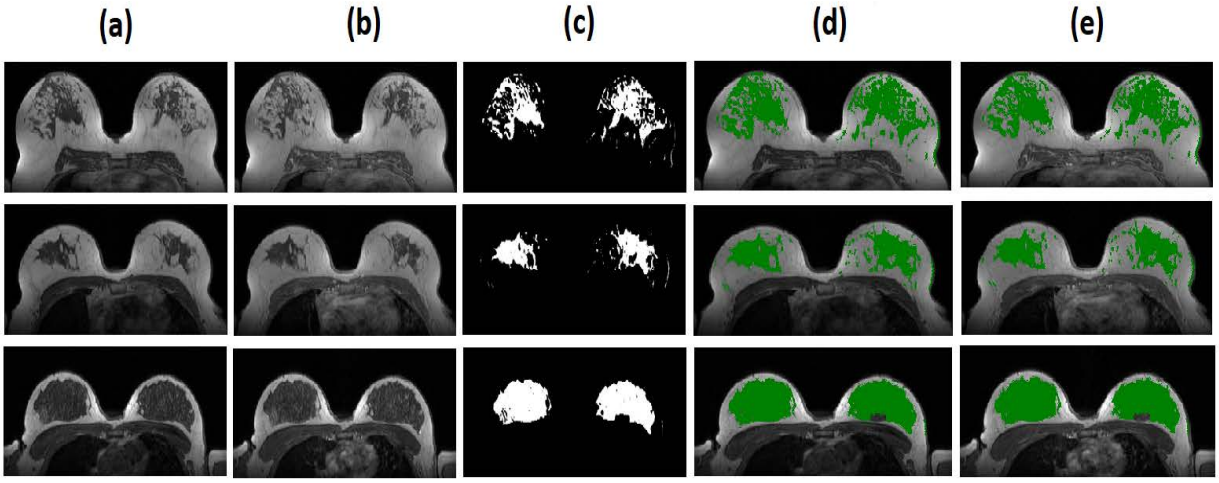


Figure 4: Axial representations of segmentated FGT of 3 cases: (a) pre-contrast (t0) image (b) first post-contrast (t1) image (c) DCE-MRI slice segmentation mask (d) mask applied to (t0) image (e) mask applied to (t1) image.

3.4. Traditional Machine Learning Approach

3.4.1. Feature Extraction

Based on a prior knowledge of the FGT and BPE volumes, enhancement (intensity) and statistical features were extracted from the FGT region. Initially, intensity information extracted from the segmented regions in both pre-contrast and post-contrast volumes was compared. Bias-field correction algorithms were discarded because it applies non-linear normalization to the volume intensities for each time point independently. Additionally, in some cases it reduced the intensity difference between pre- and post-contrast volumes. Hence, voxel information was extracted from non-bias-field corrected volumes. A number of intensity and statistical texture based features were extracted from DCE-MRI (Table 2). For the first 12 intensity-based features, relative difference between pre- and post- contrast volumes were calculated. The average distance between bins of two intensity histograms were also computed. With the aid of software library developed by Philips and Li, Haralick statistical texture features were extracted from the 3D volume. The software initially generates a 3D gray level co-occurrence matrix (GLCM) based on the number of intensity levels in the 3D volume. From the generated GLCM, several metric were measured. Table 2 is a list of all extracted features. A total of 595 features were obtained per each case of the dataset.

3.4.2. Data Augmentation

Considering the high imbalanced nature of the dataset as displayed in Figure 3, data augmentation was employed only in the train dataset. This technique was integrated in the adopted pipeline to help increase the generalization performance of the classifiers. The synthetic minority over-sampling technique (SMOTE) described by Chawla et al. (2002) was used for this purpose. The algorithm performs augmentation in the feature space rather than data space by taking least class samples and introducing synthetic examples along the line segments joining all of the k minority class nearest neighbors (Chawla et al., 2002). In the training dataset, SMOTE was adaptively applied to the least sample classes (BPE 3 and BPE 4) considering the imbalance ratio. The ratio used in augmenting BPE 3 and BPE 4 were 2 and 3 respectively. Figure 5 shows the difference before and after applying the SMOTE algorithm.

Figure 5: Histogram plot of a sample train dataset showing the impact of SMOTE. BPE 1 and BPE 2 remained constant because no augmentation was applied to these samples while BPE 3 and BPE 4 increased by a factor of 2 and 3 respectively after applying SMOTE.

Table 2: List of features extracted from DCE-MRI volumes for the traditional machine learning classification approach.

Intensity Features	Relative Difference	First quartile (Q1)
		Second quartile (Q2)
		Third quartile (Q3)
		Mean value (Mean)
		Standard deviations (std)
		Maximum value (max)
		Minimum value (min)
		Most frequent value (mode)
		Variance (var)
		Mean absolute deviation (mad)
		Skewness (skew)
		Kurtosis (kur)
	Histogram Distance	Pearson correlation coefficient
		Euclidean distance between histograms
		Chi-squared distance between histograms
		Cosine distance between histograms
		Earth moving distance (EMD) between histograms
		L1 distance between histograms
Statistical Features	Haralick	Energy
		Entropy
		Correlation
		Contrast
		Inverse Difference Moment
		Variance
		Sum Mean
		Inertia
		Cluster Shade
		Cluster Prominence
		Max Probability
		Inverse Variance
		Mode Probability

3.4.3. Classification

The final phase of the traditional machine learning approach involved training and testing the extracted features. Experiments were performed using the well-known SVM and random forest (RF) classifiers. For initial configuration of the parameters of each classifier, the dataset was randomly divided into 80% training and 20% testing. The optimal number of trees used in tuning the random forest classifier was 100 trees with batch size of 50. For classification using SVM, the linear kernel with complexity constant (C-regularizer) of 5.0 produced optimal results. Afterwards, 5 fold cross validation on the whole dataset was also carried out. That is, the whole dataset was divided in 5 subsets of similar size. One subset represents a test set (20%) and the remaining 4 as train set (80%). This procedure was repeated for all the different subset until the entire dataset was tested. Metrics used to evaluate the performance of the proposed classification tools were extracted from the corresponding 4-classes confusion matrix. These included overall accuracy (Acc), accuracy per class (Acc_{BPEn}) and $F1\_score$. Equations used in measuring the accuracy per class, overall accuracy and F1_score are displayed in Equation 2, 3 and 4 respectively. In addition, the κ was employed to quantify the agreement between the predicted labels and true labels. The significance level between predicted labels and true labels was also determined using p-value, where p-value < 0.05 represents high significance.

$$Acc_{BPEn} = \frac{TP_{BPEn}}{TP_{BPEn} + FP_{BPEn}}, \tag{2}$$

where Acc_{BPEn} is the accuracy of BPE n class, TP_{BPEn} is the true positive of BPE n class and FP_{BPEn} is the false positive of BPE n class.

$$Acc = \frac{TP}{TP + FP + TN + FN}, \tag{3}$$

where Acc is the overall accuracy, TP is the true positive, FP is the false positive, TN is the true negative and FN is the false negative.

$$F1\_score = \frac{2(pre * rec)}{pre + rec}, \tag{4}$$

where pre is the precision and rec is the recall.

3.5. Deep Learning Approach

3.5.1. Data Preparation

Unlike the traditional machine learning approach, data preparation adopted in the deep learning approach differed. This is because the performance of deep neural network is highly dependent on the input data. As such, various image preprocessing strategies were applied to the input data before feeding it to the network. Considering the different volumes due to the different MRI scanners, an isotropic voxel spacing of $(1 \times 1 \times 1)\ mm^3$ was generated for pre-contrast (t0), post-contrast (t1) and the first post-contrast subtraction (t1-t0) volumes. Subsequently, 5 slices were selected from the middle of each train volume and 1 middle slice for each test volume. For each selected slice, the corresponding t0, t1 and (t1-t0) slices were stacked together to generate 3-channel images. Using bilinear interpolation, images were rescaled to ImageNet (Deng et al., 2009) standard dimension of 224×224 pixels. A sample case is shown in Figure 6.

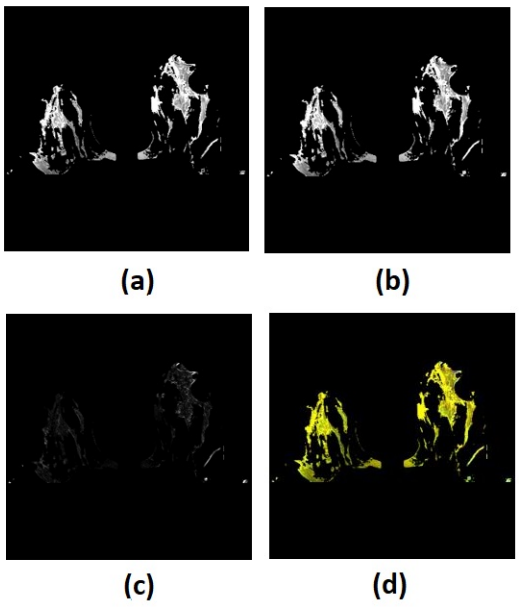
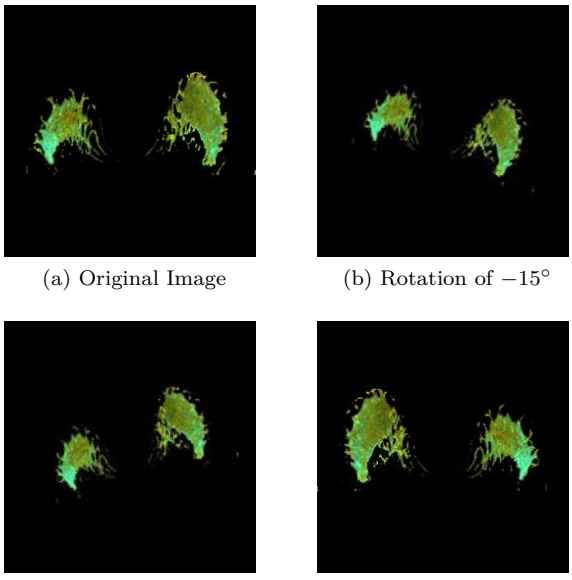


Figure 6: Data pre-processing for deep learning approach. (a) selected t0 axial image (b) corresponding t1 axial image (c) t1 - t0 difference image (d) 3-channels image. It consists of t0 image as the red channel, t1 image as green channel and t1-t0 as the blue channel.

The 3-channel images were independently normalized using mean and standard deviation of (0.485, 0.456, 0.406) and (0.229, 0.224, 0.225) respectively in accordance to ImageNet reference. The aim of this adopted normalization is to help the deep neural network to learn faster and also ensure that gradients act uniformly for each input channel (Ioffe and Szegedy, 2015).

3.5.2. Data Augmentation

Although 5 non-zero slices were used for training from each train volume, the high imbalance ratio among the 4 BPE classes still persisted. According to Perez and Wang (2017), data augmentation produces promising strategies to increase the accuracy of classification tasks and hence data augmentation was performed in the least samples to somehow balance the classes. For this reason, rotation of $\pm 15°$ and horizontal flip were applied to BPE 3 and BPE 4 of the train dataset. In contrast to the traditional machine learning approach, data augmentation was performed in the data space. An example is displayed in Figure 7.



(a) Original Image (b) Rotation of $-15°$

(c) Rotation of $+15°$ (d) Horizontal Flip

Figure 7: An illustration of data augmentation performed for smaller sample classes of the train dataset.

3.5.3. Implementation Details

Several classification networks such as Vgg-16, Vgg-19, Resnet-50 and Densenet-121 were analyzed and investigated for this tasks. However, Resnet-50 was selected over other network architectures because of its outstanding performance in the evaluation phase. This can attributed to its ability to re-use features from upper convolutional layers and also solve the vanishing gradient problem with dense networks (He et al., 2016). The baseline architectures of Resnet-50 follows the same trends as a plain network with 50 layers except that shortcut connections (residual blocks) are added to each pair of 3×3 filters. Each residual block follows the bottleneck design. That is, a stack of 3 layers consisting of 1×1, 3×3 and 1×1 convolutions, where the 1×1 layers are responsible for reducing and then restoring dimensions, leaving the 3×3 layer a bottleneck

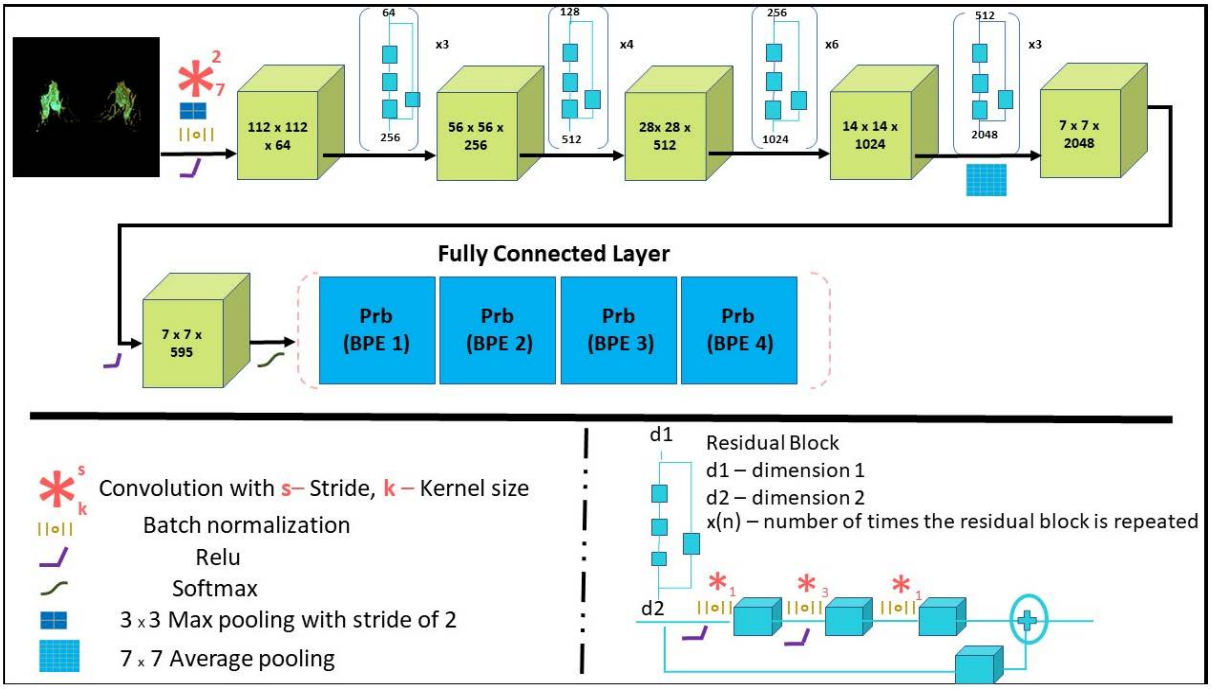


Figure 8: Overview of the proposed Resnet-50 architecture. The convolutional layers were pre-trained on ImageNet 1k natural images dataset. The convolutional feature extractors and weights of the hidden layers were frozen and transferred directly to BPE classification. Prb represents the probability.

with smaller input/output dimensions (He et al., 2016). Using the concepts of transfer learning, pre-trained Resnet-50 on ImageNet 1k natural images was used. The pre-trained model was obtained from torchvision (PyTorch) archives. Afterwards, the model was fine-tuned to fit the task at hand. In the fine-tuning phase, all the hidden layers were frozen and the last fully connected layer was modified by reducing the number of features from 2048 to 595 using ReLU weight initialization followed by a dropout layer with probability of 0.2. The role of the dropout layer was to help reduce overfitting while training the model (Srivastava et al., 2014). Hence, the new last fully connected layer was made up of 595 features classified into 4 classes using the softmax (SM) classifier. As displayed in Figure 8, the input to the network was a 3-channel image made up of 3 set of images: pre-contrast (t0) as the red channel, post-contrast (t1) as the green channel and the difference image (t1-t0) as the blue channel. Therefore, each input image represents a mini-batch of 3. For every batch size of 10, the train samples were shuffled to avoid biased class result. The cross entropy loss function (Equation 5) was used to estimate the loss after every epoch.

$$H(T,q) = -\Sigma_{i=1}^{N} \frac{1}{N} log_2 q(x_i), \qquad (5)$$

where N is the size of the test set, $q(x)$ is the probability of event, x is estimated from the training set and $H(x)$ is the estimated cross entropy loss.

This loss function was particularly useful for this task because the dataset was unbalanced. Gradients for back-propagation were estimated using stochastic gradient descent optimizer with momentum of 0.9. An initial learning rate of 0.001 was used and periodically decayed by a factor of 0.1 using a scheduler after the completion of every 10 epochs. 20% of the train set was used as validation set to investigate if the network was either overfitting or underfitting. From the 250 epochs used for training, it was observed that the optimal evaluation results were found at 100 epochs. Early stopping using the best validation accuracy as well as the least validation loss was also investigated. Just like the traditional machine learning approach, 5 five fold cross validation was performed until all the entire dataset was once used as a test set. For initial investigation and estimation of the deep neural network, the dataset was randomly divided into 80% training and 20% testing. Experiments were conducted with and without data augmentation strategies. Evaluations were performed by comparing the indices with the highest probability against the true classes using the SM classifier. Subsequently, the 4×4 confusion matrix was analyzed and various metrics such as accuracy per class, overall accuracy, F1_score, κ and p-value were computed.

		Predicted Class			
		BPE 1	BPE 2	BPE 3	BPE 4
True Class	BPE 1	23	9	7	2
	BPE 2	6	8	4	4
	BPE 3	1	1	1	0
	BPE 4	0	0	0	1

(a) RF classification without SMOTE augmentation

		Predicted Class			
		BPE 1	BPE 2	BPE 3	BPE 4
True Class	BPE 1	24	10	7	1
	BPE 2	2	6	1	2
	BPE 3	2	1	2	0
	BPE 4	1	1	2	4

(b) RF classification with SMOTE augmentation

		Predicted Class			
		BPE 1	BPE 2	BPE 3	BPE 4
True Class	BPE 1	21	13	9	2
	BPE 2	7	4	2	3
	BPE 3	1	1	1	0
	BPE 4	0	0	0	2

(c) SVM classification without SMOTE augmentation

		Predicted Class			
		BPE 1	BPE 2	BPE 3	BPE 4
True Class	BPE 1	15	10	4	0
	BPE 2	5	5	1	1
	BPE 3	7	3	3	2
	BPE 4	2	0	4	4

(d) SVM classification with SMOTE augmentation

Figure 9: Confusion matrix of BPE classification with no cross validation using the traditional machine learning approach. The diagonal of each confusion matrix coloured in green represents the correctly classified class (True positive) while plain white cells represents misclassification.

4. Results

4.1. Inter-Observer Agreement

The inter-observers agreement between all radiologist (R1, R2 and R3) participating in this annotation task ranged from moderate to fair agreement with p-value < 0.05 (except agreement between R1 and R3 with p-value $= 0.0786$). However after applying the ensembling technique, the agreement between the ensembled rates (E) and the individual rating of each reader ranged from moderate to substantial with high significance p-value for each case. The inter-observers agreement and the corresponding p-value of each pair is summarized in Table 3.

Table 3: Inter-observers agreement.

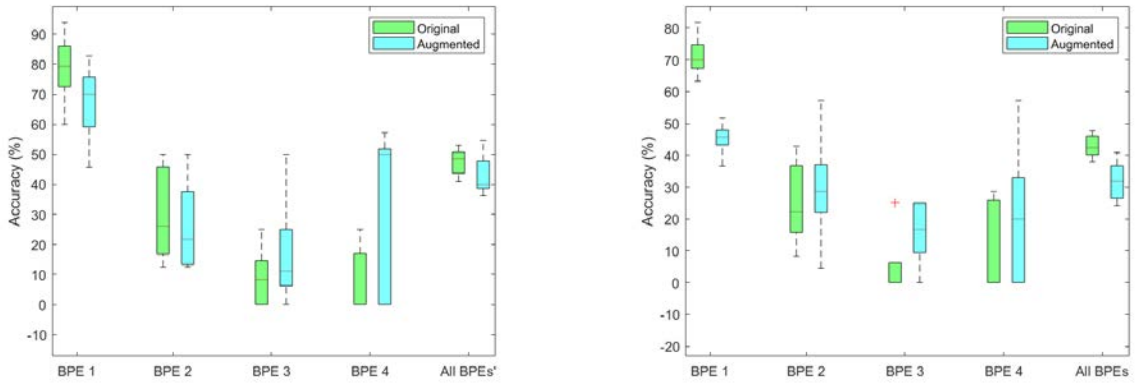
Readers' Pair	kappa (κ)	p-value	Agreement
R1 & R2	0.41	0.0004	Moderate
R1 & R3	0.18	0.0786	Slight
R2 & R3	0.13	0.0291	Slight
R1 & E	0.65	0.0000	Substantial
R2 & E	0.66	0.0000	Substantial
R3 & E	0.46	0.0000	Moderate

4.2. Traditional Machine Learning

Following the methodologies described in section 3, the performance of RF and SVM classifiers using all 595 extracted features from the segmented parenchymal volume are displayed in Figure 9. A total of 66 test cases were used to evaluate the trained models. For each experiment, cases with and without data augmentation strategies were compared. The various evaluation metrics measured from each confusion matrix is displayed in Table 4. Comparatively, the overall performance of the RF classifier was better than the SVM classifier. In Table 4, it can be observed that BPE 1 has the highest accuracy with BPE 3 having the least accuracy. The overall accuracy on the test samples was 50% with F1_score of 0.46. The kappa agreement between the true class and predicted class was computed as 0.27 with p-value $= 0.2601 > 0.05$. Hence the hypothesis of this test can be considered as insignificant. Nonetheless, after applying the SMOTE data augmentation to the least class samples .i.e: BPE 3 and BPE 4, the accuracy of each class increased significantly for all cases with the exception of BPE 2 which decreased by 10%. Additionally, other metrics such as the overall accuracy, kappa agreement and F1_score also increased accordingly. The p-value value of the true class against the predicted class changed from insignificant hypothesis to significant hypothesis (p-value $= 0.0034 < 0.05$). Relatively, the changes observed in the SVM classifier after applying the SMOTE algorithm followed a similar trend as the RF classifier. Although the accuracy per class increased after SMOTE in all the classes (except of BPE 1), the overall accuracy and F1_score without SMOTE augmentation was slightly higher (Overall accuracy $= 42.42\%$ without SMOTE and 40% with SMOTE, F1_score $= 0.40$ without SMOTE and 0.39 with SMOTE. However, the κ agreement between the

Classifier	Accuracy per class (%)				Accuracy (%)	F1-score	kappa (κ)	p-value
	BPE 1	BPE 2	BPE 3	BPE 4				
RF	79.31	44.44	8.33	14.29	50.00	0.46	0.27	0.2601
RF*	82.76	33.33	16.67	57.14	54.55	0.49	0.44	0.0034
SVM	72.41	22.22	8.33	28.57	42.42	0.40	0.23	0.3160
SVM*	51.72	27.78	25.00	57.14	40.91	0.39	0.41	0.0016

Table 4: Results of evaluation metrics from Figure 9. * represents results for a more balanced training dataset using SMOTE.



(a) RF (b) SVM

Figure 10: Boxplot of 5 fold cross validation using traditional machine learning approach. The green box represents the original dataset without SMOTE augmentation while the blue box represents dataset with SMOTE augmentation. The first 4 pairs of box and whisker plots show distribution of accuracy as a function of each BPE class assessment. The last pair represents the overall accuracy for all the 5 fold. The red '+' symbol represents an outlier.

predicted and true class was significantly higher by a factor of 2 after applying the SMOTE algorithm. As previously observed with the RF classifier, the p-value measured between the true class against the predicted class also changed from insignificant hypothesis to significant hypothesis (p-value = 0.0016 $<$ 0.05). Afterwards, the same set of evaluation were performed on the other fold until all the cases were once used as a test case. The results obtained for the 5 folds are summarized in the boxplot shown in Figure 10. Each box indicates the median accuracy as line dividing the box with the interquartile range as the height of the box. The whiskers extend from both side of the box for most of cases. This shows the ranges for the bottom 25% and the top 25% of the data values, excluding outliers. With regards to the metrics measured for the 5 folds, it is evident as displayed in Figure 10 that the overall accuracy (Acc) was compact for both classifiers than the individual classes especially cases without the SMOTE augmentation. For experiments conducted with the RF classifier, none of the accuracy measured for all 5 folds was recorded as an outlier in the boxplot. However, the SVM classifier recorded 1 outlier case as illustrated in Figure10b. This was an extreme case when added to the boxplot could have a huge on the impact on the final appearance.

4.3. Deep Learning

In order to adequately compare the results of the 2 adopted methods (i.e: traditional machine learning and deep learning), the same train and test cases used in the traditional machine learning approach were used for the deep learning experiments. A total of 66 test cases were used to validate the trained models. The results obtained for each case study is displayed in Figure 11. Table 5 shows the various metrics measured from the 4×4 confusion matrix. From Table 5, it can be deduced that BPE 4 has the highest accuracy with (Acc_{BPE4} = 75%) while BPE 2 has the least accuracy with (Acc_{BPE2} = 50%) when no data augmentation strategies were applied. For the same experiment, the computed overall accuracy was 58% with $F1\_score$ and κ of 0.55 and 0.40 respectively. The p-value between the true class and predicted class was 0.0170 $<$ 0.05, thus the hypothesis of this test was highly significant. However, after data augmentation was applied to the least class samples such as BPE 3 and BPE 4, there was a drastic reduction in all the measured metrics with the exception of Acc_{BPE4}. The accuracy of BPE 4 with and without augmentation was the same for both cases. Subsequently, 5 fold cross validation was performed. The results obtained for the acccuracy per class (Acc_{BPEn}) and

		Predicted Class			
		BPE 1	BPE 2	BPE 3	BPE 4
True Class	BPE 1	23	3	3	0
	BPE 2	10	7	0	1
	BPE 3	4	3	5	0
	BPE 4	3	1	0	3

(a) Classification without data augmentation

		Predicted Class			
		BPE 1	BPE 2	BPE 3	BPE 4
True Class	BPE 1	24	5	0	0
	BPE 2	11	5	2	0
	BPE 3	7	2	2	1
	BPE 4	3	1	0	3

(b) Classification with data augmentation

Figure 11: Confusion matrix of BPE classification with no cross validation using deep learning approach. The diagonal of each confusion matrix coloured in green represents the correctly classified class (True positive) while plain white cells represents misclassification.

Classifier	Accuracy per class (%)				Accuracy (%)	F1-score	kappa (κ)	p-value
	BPE 1	BPE 2	BPE 3	BPE 4				
SM	57.50	50.00	62.50	75.00	58.00	0.55	0.40	0.0170
SM*	53.33	38.46	50.00	75.00	51.00	0.48	0.39	0.0528

Table 5: Results of evaluation metrics from Figure 11. * represents results for a more balanced training dataset using data augmentation strategies. SM represents Softmax classifier.

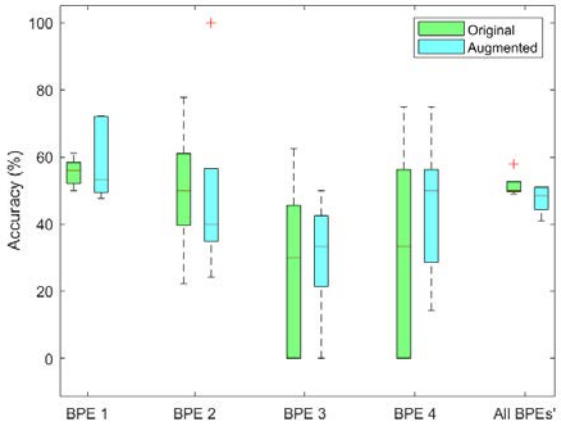


Figure 12: Boxplot of 5 fold cross validation using deep learning approach. The green box represents the original datasets without data augmentation while the blue box represent datasets with data augmentation. The first 4 pairs of box and whisker plots show distribution of accuracy as a function of each BPE class assessment. The last pair represents the overall accuracy for all the 5 fold. The red '+' symbol represents an outlier.

overall class is summarized in the boxplot displayed in Figure 12. As illustrated in Figure 12, the accuracy of BPE 1 (Acc_{BPE1}) without any data augmentation strategies was more compact than the other BPE classes. However with data augmentation strategies, the accuracy of BPE 3 (Acc_{BPE3}) was more compact than the other classes. Relatively, the overall accuracy without any data augmentation strategies was also more compact than data augmentation strategies. This implies that for all folds, the accuracies measured at each fold were similar relative to each other with less deviation. 2 outliers were recorded as displayed Figure 12.

5. Discussion

This study presents machine learning techniques (traditional machine learning and deep learning) for automatic BPE level classification following the 4 ordinal categories defined by ACR. Although these 4 BPE categories are clearly described by the ACR, manual annotation currently undertaken by radiologists suffers from large intra- and inter-observer variability as proven by the inter-observer's agreement results obtained for the 3 readers ranging from slight to moderate agreement as shown in Table 3. As displayed in Table 1, the large variability observed among the 3 readers originates from their level of experience, speciality and training. Agreement between R1 and R2 was moderate with κ = 0.41 and p-value of 0.0004 $<$ 0.05. However, agreement between R1 and R3 was slight with κ = 0.13 and p-value of 0.0291 $<$ 0.05. Likewise, the agreement between R2 and R3 with κ = 0.18 and p-value of 0.0786 $>$ 0.05. This large variability was mostly caused by R3 because of his field of speciality (prostate). Aside the large intra- and inter-observer variability observed in current BPE evaluation by radiologists, they are very tedious and time-consuming tasks. These discrepancies might be regarded as an initial drawback as previously observed in the literature by Dontchos et al. (2015) and (Pujara et al., 2018). Hence, there is the need for a further (quantitative and

automated) re-definition of the BPE rating criteria. These automated BPE classification tools, as proposed in this study has the potential to reduce the tedious, subjective nature of BPE classification process. Additionally, such tools could help in training young radiologists in BPE classification or even the reduction of inter- and intra-reader variability. Up to date, many literatures on quantitative BPE assessment have been performed by Klifa et al. (2011), Yang et al. (2015), Ha et al. (2016) and Pujara et al. (2018). Nevertheless, their methods are still time-consuming and require initial delineation of the ROI i.e. FGT by the radiologist which may introduce potential subjective bias (Eyal et al., 2009)(Clendenen et al., 2013)(Ha et al., 2019). Due to these limitations posed by the semi-automated approaches, fully automated techniques are needed to completely remove the subjective nature of BPE classification. Ha et al. (2019) recently developed a fully automated approach based on the convolutional neural network (CNN) for FGT quantification and BPE classification. For further validation with the task started by Ha et al. (2019), the suggested approach in this study also adopts automated techniques based on machine learning algorithms (traditional machine learning and deep learning).

In order to establish ground truth labels for the automatic approaches, the individual rates from the 3 radiologists were fused together using the majority voting ensembling technique. That is, for each study, the final label was assigned based on the agreement between at least 2 readers. Prior to this ensembling technique, cases without any agreement with at least 2 readers were also filtered out. That is, patient study where the 3 readers had 3 independent rate (R1 - BPE 1, R2 - BPE 4 and R3 - BPE 2). In this scenerio, the mean of the 3 rate gives a value of 2.6 which is ≈ 3. However, none of the readers evaluated the rate as BPE 3. Therefore, to avoid wrong label assignment, cases without agreement between at least 2 readers were filtered out. Subsequently, the agreement between the ensembled rates and each individual readers' rate were determined and found to range from moderate to substantial agreement. These observed agreement is indicative that all the individual rates were fairly considered while assigning the final labels.

Although the goal of this study is a classification task, the performance of the suggested automated techniques are highly dependent on the correct segmentation of the ROI i.e: FGT. For these reasons, robust and fully automated segmentation of the whole breast and FGT segmentation developed by Gubern-Merida et al. (2014) were applied to each case study independently to generate segmentation mask. Segmentation masks were generated only in (t0) volumes and used for all the time points. This is because there were no records of patient motion between the different time points. As stated previously, the segmentation obtained in some cases were suboptimal and hence some of the segmentation masks were manually corrected.

With respect to the use of the traditional machine learning approach in conjunction with the the hand crafted (enhancement and statistical haralick texture) features, cases with and without data augmentation strategies were compared in each classifiers. For initial estimation of the performance of each classifier, the dataset was divided into 80% training made of 263 cases and 20% testing made up of 66 cases. Comparatively, the overall performance of the RF classifier was better than the SVM classifier. This can be associated its ability to reduce risk of overfitting. Additionally, it is good tool for multiclass problem. From the results obtained as displayed Table 4, it can be observed that BPE 1 recorded the highest accuracy for both RF and SVM classifiers. The high accuracy observed in BPE 1 class can be attributed to many sample cases used in training the model as displayed in Figure 5. Nevertheless, after applying the SMOTE data augmentation strategies to the least class samples (BPE 3 and BPE 4), there was a significant increase in all the individual BPE classes (excluding BPE 2 which reduced by 10%). Surprisingly, applying the SMOTE data augmentation algorithm to the least sample classes positively affected the p-value between the predicted and true class for each classifier by changing its value from insignificant hypothesis to significant hypothesis. That is, the p-value of the true class against the predicted class without SMOTE data augmentation was measured as 0.2601 which exceed the threshold value of 0.05 (insignificant) and p-value of the true class against the predicted class with SMOTE was measured as 0.0014 which within the threshold value of 0.05 (significant). From the 5 fold cross validation carried out, results obtained (Figure 10) shows that the overall accuracy (Acc) was compact for both classifiers than the individual classes especially cases without the SMOTE augmentation. This implies that the overall accuracy measured for all 5 folds were less deviated from each other when compared to the individual class cases.

The deep learning algorithms were also used to investigate BPE classification. Unlike the traditional machine learning approach, features were automatically extracted by the deep neural network. The input to the network were set of 2D slices selected from the middle volume of each case study. In the data preparation phase, the t0, t1 and (t1-t0) images were stacked together to generate 3-channel images. This step was undertaken to ensure that

Table 6: Comparison between published works in section 2 on BPE classification and our adopted approach. * represents results for a more balanced training dataset using data augmentation strategies. SM represents Softmax classifier. **Bold** numbers represent the maximum value along the column. '-' represents unknown metrics.

Reference	$AccBPE_1$ (%)	$AccBPE_2$ (%)	$AccBPE_3$ (%)	$AccBPE_4$ (%)	Acc (%)	AUC
Klifa et al. (2011)	23.00	22.00	17.00	23.00	-	-
Yang et al. (2015)	-	-	-	-	-	0.865
Ha et al. (2016)	4.61	8.74	18.10	37.40	-	-
Pujara et al. (2018)	20.20	25.20	50.00	50.00	-	-
Ha et al. (2019)	-	-	-	-	**82.90**	-
Our Approach (SVM)	72.41	22.22	8.33	28.57	42.42	-
Our Approach (SVM^*)	51.72	27.78	25.00	57.14	40.91	-
Our Approach (RF)	79.31	44.44	8.33	14.29	50.00	-
Our Approach (RF^*)	**82.76**	33.33	16.67	57.14	54.55	-
Our Approach (SM)	57.50	**50.00**	**62.50**	**75.00**	58.00	-
Our Approach (SM^*)	53.33	38.46	50.00	**75.00**	51.00	-

the deep neural network learns effectively because most of the subtraction (t1-t0) images were negligible. That is, the sum of their intensity were 0. Using the same train and test cases, the deep neural network were trained using pretrained Resnet-50 model. In contrast to the traditional machine learning algorithm, this approach was highly dependent on the data size and data preprocessing. Initially 1 middle slice was used for both training and testing. However, the generalization results obtained after evaluating the trained model was suboptimal and as such for each train samples, instead of 1 middle slice, 5 middle slices were selected. In addition, transfer learning was also considered over training from scratch. Another alternative considered in the data preparation phase was projection of all the 2D slices into one projected slice. This data preparation technique resulted in pixelated images and hence their use were discarded. Data augmentation strategies were introduced such as rotation of $\pm$ 15° and horizontal flip as displayed in Figure 7 were applied to the least class samples in attempt balance the classes. Contrary to the data preparation adopted in the training phase, only 1 middle slice was used to evaluate the trained model during the validation phase. This was done to make the deep learning algorithm more comparable to the traditional machine learning approach. The hyperparameters such as the batch size, learning rate and epoch in the deep neural networks were carefully tuned to obtained optimal results. To investigate whether the network was either overfitting or underfitting, 20% of the train set were used as validation set to monitor the trends of both validation loss with respect to the train loss. The deviation between the train and validation loss was less for first 100 epochs covered. However, the validation loss increase progressively afterwards. Early stopping using best validation accuracy and least validation loss were investigated. This technique saves the model with best validation accuracy. Their use was later discarded because, for most experiments the results obtained without early stopping was much better when compared to the early stopping techniques. Hence, fixed epochs were set and used throughout all the experiments.

Comparatively, the results obtained for the deep learning approach were slightly higher than that of the traditional machine learning approach. That is, the best overall accuracy measured using the traditional machine approach without data augmentation was computed as 50% while that of the deep learning was found to be 58%. As displayed Table 5, all the 4 BPE classes received fair accuracy per class ($AccBPE_n$), hence the accuracy per class was not biased towards classes with many samples as it was the case for the traditional machine learning approach. As illustrated Figure 12, the accuracy per class ($AccBPE_n$) for the SM classifier were relatively low when compared to RF and SVM classifiers. These variations were as a result of the random division employed while splitting the data set into 5 folds. Hence folds with well balanced classes performed better than folds with huge imbalance among the classes. Unlike the traditional machine learning approach, the adopted data augmentation in the deep learning method did not improve any of the metrics measured as observed on Table 5.

In the end, we compared our results to already published works on BPE level classification as shown in Table 6. From the table, our approach received the best accuracies in terms of accuracy per class ($AccBPE_n$). However, Ha et al. (2019) received the best overall accuracy (Acc). The high performance observed in their approach can be linked to the good baseline used to establish ground truth as well as the robust segmentation tool used in segmenting FGT prior to the classification tasks. Nevertheless, our approach provides a good baseline for further evaluation.

6. Conclusions

Machine learning algorithms (traditional machine learning and deep learning) were investigated in this study as a supplementary tool for radiologists in BPE level classification. The classification models described in this work represents a step towards an automated and objective classification of BPE level following ACS recommendations. The main advantages of such tools is the reduction of inter- and intra-reader variability observed during manual BPE classification. Such automated tools could also be beneficial for radiologists with less experience as they could be trained on the annotations of a more experienced radiologists. Furthermore, automatic BPE classification tools could be included into other applications such as breast cancer risk estimation models for patient stratification and risk assessment in personalised screening scenarios.

For future studies, more BPE readers should be engaged in rating BPE levels since the suggested techniques are dependent on the manual annotations given by radiologists'. Additionally, the 3D U-Net architecture can be used to initially segment the ROI before passing it to the various suggested algorithms. Furthermore, larger dataset will likely improve our models. Also, the use of N4 bias field correction could further be explored.

7. Acknowledgments

I would like to thank my supervisors; Dr. Oliver Diaz and Dr. Robert Martí for their constant support, guidance, useful suggestions and encouragement throughout this project. I am very grateful.

References

Baltzer, P., Dietzel, M., Vag, T., Burmeister, H., Gajda, M., Camara, O., Pfleiderer, S., Kaiser, W., 2011. Clinical mr mammography: impact of hormonal status on background enhancement and diagnostic accuracy, in: RöFo-Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren, © Georg Thieme Verlag KG Stuttgart· New York. pp. 441–447.

Bennani-Baiti, B., Dietzel, M., Baltzer, P.A., 2016. Mri background parenchymal enhancement is not associated with breast cancer. PloS one 11, e0158573.

Braman, N.M., Etesami, M., Prasanna, P., Dubchuk, C., Gilmore, H., Tiwari, P., Plecha, D., Madabhushi, A., 2017. Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast dce-mri. Breast Cancer Research 19, 57.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research 16, 321–357.

Clendenen, T.V., Zeleniuch-Jacquotte, A., Moy, L., Pike, M.C., Rusinek, H., Kim, S., 2013. Comparison of 3-point dixon imaging and fuzzy c-means clustering methods for breast density measurement. Journal of Magnetic Resonance Imaging 38, 474–481.

DeMartini, W.B., Liu, F., Peacock, S., Eby, P.R., Gutierrez, R.L., Lehman, C.D., 2012. Background parenchymal enhancement on breast mri: impact on diagnostic performance. American Journal of Roentgenology 198, W373–W380.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee. pp. 248–255.

Dontchos, B.N., Rahbar, H., Partridge, S.C., Korde, L.A., Lam, D.L., Scheel, J.R., Peacock, S., Lehman, C.D., 2015. Are qualitative assessments of background parenchymal enhancement, amount of fibroglandular tissue on mr images, and mammographic density associated with breast cancer risk? Radiology 276, 371–380.

Eyal, E., Badikhi, D., Furman-Haran, E., Kelcz, F., Kirshenbaum, K.J., Degani, H., 2009. Principal component analysis of breast dce-mri adjusted with a model-based method. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine 30, 989–998.

Felix, A.S., Lenz, P., Pfeiffer, R.M., Hewitt, S.M., Morris, J., Patel, D.A., Geller, B., Vacek, P.M., Weaver, D.L., Chicoine, R.E., et al., 2016. Relationships between mammographic density, tissue microvessel density, and breast biopsy diagnosis. Breast Cancer Research 18, 88.

Giess, C.S., Yeh, E.D., Raza, S., Birdwell, R.L., 2014. Background parenchymal enhancement at breast mr imaging: normal patterns, diagnostic challenges, and potential for false-positive and false-negative interpretation. Radiographics 34, 234–247.

Gubern-Merida, A., Kallenberg, M., Mann, R.M., Marti, R., Karssemeijer, N., 2014. Breast segmentation and density estimation in breast mri: a fully automatic framework. ieee journal of biomedical and health informatics 19, 349–357.

Gubern-Mérida, A., Martí, R., Melendez, J., Hauth, J.L., Mann, R.M., Karssemeijer, N., Platel, B., 2015. Automated localization of breast cancer in dce-mri. Medical image analysis 20, 265–274.

Ha, R., Chang, P., Mema, E., Mutasa, S., Karcich, J., Wynn, R.T., Liu, M.Z., Jambawalikar, S., 2019. Fully automated convolutional neural network method for quantification of breast mri fibroglandular tissue and background parenchymal enhancement. Journal of digital imaging 32, 141–147.

Ha, R., Mema, E., Guo, X., Mango, V., Desperito, E., Ha, J., Wynn, R., Zhao, B., 2016. Three-dimensional quantitative validation of breast magnetic resonance imaging background parenchymal enhancement assessments. Current problems in diagnostic radiology 45, 297–303.

Hambly, N.M., Liberman, L., Dershaw, D.D., Brennan, S., Morris, E.A., 2011. Background parenchymal enhancement on baseline screening breast mri: impact on biopsy rate and short-interval follow-up. American Journal of Roentgenology 196, 218–224.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Hruska, C.B., Scott, C.G., Conners, A.L., Whaley, D.H., Rhodes, D.J., Carter, R.E., O'Connor, M.K., Hunt, K.N., Brandt, K.R., Vachon, C.M., 2016. Background parenchymal uptake on molecular breast imaging as a breast cancer risk factor: a case-control study. Breast Cancer Research 18, 42.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 .

Kamińska, M., Ciszewski, T., Łopacka-Szatan, K., Miotła, P., Starosławska, E., 2015. Breast cancer risk factors. Przeglad menopauzalny= Menopause review 14, 196.

King, V., Brooks, J.D., Bernstein, J.L., Reiner, A.S., Pike, M.C., Morris, E.A., 2011. Background parenchymal enhancement at breast mr imaging and breast cancer risk. Radiology 260, 50–60.

King, V., Goldfarb, S.B., Brooks, J.D., Sung, J.S., Nulsen, B.F., Jozefara, J.E., Pike, M.C., Dickler, M.N., Morris, E.A., 2012. Effect of aromatase inhibitors on background parenchymal enhancement and amount of fibroglandular tissue at breast mr imaging. Radiology 264, 670–678.

Klifa, C., Suzuki, S., Aliu, S., Singer, L., Wilmes, L., Newitt, D., Joe, B., Hylton, N., 2011. Quantification of background enhancement in breast magnetic resonance imaging. Journal of Magnetic Resonance Imaging 33, 1229–1234.

Kuhl, C.K., Mielcareck, P., Klaschik, S., Leutner, C., Wardelmann, E., Gieseke, J., Schild, H.H., 1999. Dynamic breast mr imaging: are signal intensity time course data useful for differential diagnosis of enhancing lesions? Radiology 211, 101–110.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. biometrics , 159–174.

Melsaether, A., Pujara, A.C., Elias, K., Pysarenko, K., Gudi, A., Dodelzon, K., Babb, J.S., Gao, Y., Moy, L., 2017. Background parenchymal enhancement over exam time in patients with and without breast cancer. Journal of Magnetic Resonance Imaging 45, 74–83.

Mema, E., Mango, V.L., Guo, X., Karcich, J., Yeh, R., Wynn, R.T., Zhao, B., Ha, R.S., 2018. Does breast mri background parenchymal enhancement indicate metabolic activity? qualitative and 3d quantitative computer imaging analysis. Journal of Magnetic Resonance Imaging 47, 753–759.

Morris, E.A., 2007. Diagnostic breast mr imaging: current status and future directions. Radiologic clinics of North America 45, 863–880.

Müller-Schimpfle, M., Ohmenhaüser, K., Stoll, P., Dietz, K., Claussen, C.D., 1997. Menstrual cycle and age: influence on parenchymal contrast medium enhancement in mr imaging of the breast. Radiology 203, 145–149.

Park, C.S., Lee, J.H., Yim, H.W., Kang, B.J., Kim, H.S., Jung, J.I., Jung, N.Y., Kim, S.H., 2007. Observer agreement using the acr breast imaging reporting and data system (bi-rads)-ultrasound, (2003). Korean journal of radiology 8, 397–402.

Perez, L., Wang, J., 2017. The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621 .

Pfleiderer, S.O., Sachse, S., Sauner, D., Marx, C., Malich, A., Wurdinger, S., Kaiser, W.A., 2004. Changes in magnetic resonance mammography due to hormone replacement therapy. Breast cancer research 6, R232.

Philips, C., Li, D., . 3d statistical texture algortihm. `https://nl.mathworks.com/matlabcentral/fileexchange/19058-cooc3d`. [Online; accessed 18-May-2019].

Pujara, A.C., Mikheev, A., Rusinek, H., Gao, Y., Chhor, C., Pysarenko, K., Rallapalli, H., Walczyk, J., Moccaldi, M., Babb, J.S., et al., 2018. Comparison between qualitative and quantitative assessment of background parenchymal enhancement on breast mri. Journal of Magnetic Resonance Imaging 47, 1685–1691.

Schrading, S., Schild, H., Kühr, M., Kuhl, C., 2014. Effects of tamoxifen and aromatase inhibitors on breast tissue enhancement in dynamic contrast–enhanced breast mr imaging: a longitudinal intraindividual cohort study. Radiology 271, 45–55.

Siegel, R.L., Miller, K.D., Jemal, A., 2019. Cancer statistics, 2019. CA: a cancer journal for clinicians 69, 7–34.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research 15, 1929–1958.

Uematsu, T., Kasami, M., Watanabe, J., 2011. Does the degree of background enhancement in breast mri affect the detection and staging of breast cancer? European radiology 21, 2261–2267.

van der Velden, B.H., Sutton, E.J., Carbonaro, L.A., Pijnappel, R.M., Morris, E.A., Gilhuijs, K.G., 2018. Contralateral parenchymal enhancement on dynamic contrast-enhanced mri reproduces as a biomarker of survival in er-positive/her2-negative breast cancer patients. European radiology 28, 4705–4716.

Vignati, A., Giannini, V., De Luca, M., Morra, L., Persano, D., Carbonaro, L.A., Bertotto, I., Martincich, L., Regge, D., Bert, A., et al., 2011. Performance of a fully automatic lesion detection system for breast dce-mri. Journal of Magnetic Resonance Imaging 34, 1341–1351.

Vreemann, S., Dalmis, M.U., Bult, P., Karssemeijer, N., Broeders, M.J., Gubern-Mérida, A., Mann, R.M., 2019. Amount of fibroglandular tissue fgt and background parenchymal enhancement bpe in relation to breast cancer risk and false positives in a breast mri screening program. European radiology , 1–13.

Vreemann, S., Gubern-Mérida, A., Borelli, C., Bult, P., Karssemeijer, N., Mann, R.M., 2018. The correlation of background parenchymal enhancement in the contralateral breast with patient and tumor characteristics of mri-screen detected breast cancers. PloS one 13, e0191399.

Wu, S., Berg, W.A., Zuley, M.L., Kurland, B.F., Jankowitz, R.C., Nishikawa, R., Gur, D., Sumkin, J.H., 2016. Breast mri contrast enhancement kinetics of normal parenchyma correlate with presence of breast cancer. Breast Cancer Research 18, 76.

Yang, Q., Li, L., Zhang, J., Shao, G., Zheng, B., 2015. A new quantitative image analysis method for improving breast cancer diagnosis using dce-mri examinations. Medical physics 42, 103–109.

You, C., Kaiser, A.K., Baltzer, P., Krammer, J., Gu, Y., Peng, W., Schönberg, S.O., Kaiser, C.G., 2018. The assessment of background parenchymal enhancement (bpe) in a high-risk population: What causes bpe? Translational oncology 11, 243–249.

Medical Imaging and Applications

Master Thesis, June 2019

Active Learning: Smart Sample Selection for Efficient Medical Image Annotation

Daria Zotova, Aneta Lisowska

Canon Medical Research Europe Ltd., Edinburgh, United Kingdom

Abstract

Supervised machine learning techniques require large amounts of annotated training data to attain good performance. However, annotated data are difficult and expensive to obtain, especially in the medical domain where only domain experts, whose time is scarce, can provide reliable labels. Active learning aims to ease the data collection process by automatically deciding which instances an expert should annotate in order to train a model as quickly and effectively as possible. In this study we evaluate different data selection approaches (random, uncertain, and representative sampling) and a semi-supervised model training procedure (pseudo-labeling), in the context of lung nodule segmentation in CT volumes. Results showed that a strategy based on estimation of the uncertainty level and representativeness slightly outperformed the pseudo-labeling technique, but both managed to reach the highest performance with ≈ 58% of the training data.

Keywords: MAIA master, active learning, deep learning, lung nodule segmentation, annotation of biomedical data

1. Introduction

Supervised machine learning techniques require large amounts of annotated training data to attain good performance. However, annotated data are difficult and expensive to obtain, especially in the medical domain where only domain experts, whose time is scarce, can provide reliable labels. Active Learning (AL) aims to ease the data collection process by automatically deciding which instances an annotator should label to train a model as quickly and effectively as possible.

In a typical AL scenario (Figure 1) we start with a small set of of annotated data (labeled pool), which are used for the initial model training. Then an active learning strategy, which usually relies on the model predictions, selects few data samples from the unlabeled pool of data that should be annotated by a human or a machine. After chosen samples have been labeled, they are added to the labeled dataset, the model is trained again, and the whole process is repeated until the satisfying level of a model performance is reached. [1]

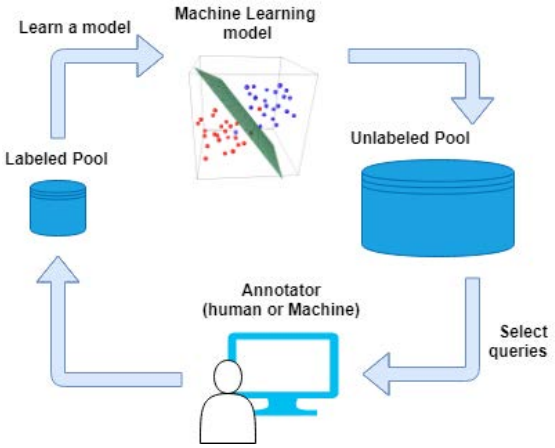


Figure 1: Active Learning procedure

The aim of this study is to implement and compare active learning strategies that have shown promising results for different tasks, and apply it within medical domain, specifically, we chose lung nodules segmentation using CNN as a primary goal.

[1] Note that to be able to assess changes in the model performance a separate set of annotated data is.

2. State of the art

In recent years a range of active learning strategies have been proposed, mostly with primary focus on classification tasks. In *'Learning Active Learning'* (LAL) by Konyushkova et al. (2017) a regression problem was formulated: given a trained classifier and its output for a specific sample without a label, the reduction in generalization error that can be expected by adding the label to that point is predicted. Two strategies were proposed for the samples selection: LAL independent, where the initial dataset is split into labeled and unlabeled randomly, and LAL iterative, where AL procedure which selects data points according to the strategy learned on the previously collected data was simulated for partitioning into labeled and unlabeled sets, and samples at each iteration depend on the samples at the previous iteration. LAL has proved to work well on real data from several different domains such as biomedical imaging, economics, molecular biology and high energy physics. One of the successful and simple strategies is called Uncertainty sampling (US). It focuses on selecting samples which the current classifier is the least certain about. The most common options to estimate the level of uncertainity:

1. Least confidence proposed by Settles (2009): all the unlabeled samples should be ranked in an ascending order according to the value of lc_i:

$$lc_i = maxp(y_i = j|x_i; W) \tag{1}$$

   where $p(y_i = j|x_i; W)$ denotes the probability of x_i belonging to the class j.
2. Margin sampling proposed by Scheffer et al. (2001): all the unlabeled samples should be ranked in an ascending order according to the value of ms_i:

$$ms_i = p(y_i = j_1|x_i; W) - p(y_i = j_2|x_i; W) \tag{2}$$

   where j_1 and j_2 the most probable classes for the sample.
3. Entropy proposed by Shannon (2001): all the unlabeled samples should be ranked in a descending order according to the value of en_i:

$$en_i = -\sum_{j=1}^{m} p(y_i = j|x_i; W) log p(y_i = j|x_i; W) \tag{3}$$

In addition to estimation of the uncertainty level in CEAL by Wang et al. (2016) a strategy that automatically selects and annotates the high confidence samples was proposed. Not only uncertain samples are added into the training set, but also the majority samples with high prediction confidence. For these certain kind of samples pseudo-labels are assigned automatically without human labour cost. From the unlabeled pool of data high confidence samples are selected as those whose entropy is smaller than a certain threshold δ, and the pseudo-label y_i is defined as follows:

$$j^* = argmaxp(y_i = j|x_i; W), \tag{4}$$

$$y_i = \begin{cases} j^*, en_i < \delta \\ 0, otherwise \end{cases} \tag{5}$$

The proposed strategy was tested on two datasets, namely Cross-Age Celebrity face recognition Dataset (CACD) and the Caltech-256 object categorization for the classification task. CEAL algorithm with margin sampling criterion outperformed other methods and reduced the need of labeled samples (63% of all labeled data for CACD were required, 76% for Caltech-256).

Gorriz et al. (2017) contributes to the previous work and apply the CEAL algorithm for medical imaging segmentation using CNN. They propose to use the effect of the dropout layer for the image uncertainty estimation. The dropout works by randomly deactivating network activations, and at test time it allows to estimate pixel-wise uncertainty. Images were taken from ISIC 2017 challenge and were split randomly into a test set with 400 images, initial labeled set with 600 images for training and the remaning 1000 images were put into the unlabeled set. At each active learning loop 10 samples with no melanoma detection, the 10 most uncertain samples and 15 random samples were added into the labeled pool for further training. A Dice score of 0.74 was reached after 9 active learning iterations with the help of 32% of data from unlabeled pool.

In Suggestive Annotation framework by Yang et al. (2017) a deep active learning framework was introduced and the goal was to determine the most representative and uncertain samples. While uncertainty means that annotated areas should be difficult for the network to perform a segmentation task, representativeness means that annotated areas need to have useful characteristics (features) for as many samples from the unlabeled pool of data as possible.

The key idea of the representativeness estimation is to define the similarity between two images. If we take the output of the last convolution layer from the encoder then it can be seen as high level features I_i^f of the image I_i. The similarity can be further estimated as $sim(I_i, I_j) = cosine\_similarity(I_i^f, I_j^f)$. If we denote S_a as an annotated set, S_u - unannotated set, then first we can define the representativeness of S_a for an image $I_x \in S_u$ as $f(S_a, I_x) = max_{I_i \in S_a} sim(I_i, I_x)$, where $sim(\cdot, \cdot)$ is the similarity measure between two images I_i and I_x. And the final step is to choose those candidates from S_a that maximize $F(S_a, S_u) = \sum_{I_j \in S_u} f(S_a, I_j)$.

The method was applied to the 2015 MICCAI Gland Challenge dataset and a lymph node ultrasound image

segmentation dataset. By using only 50% of training data state-of-the-art segmentation performance was achieved.

Ozdemir et al. (2018) proposed an active learning algorithm that was applied for gland segmentation. It followed the idea of Suggestive Annotation with two novelties for selecting samples at every active learning loop:

1. Deep contour-aware network for the segmentation was modified by adding the abstraction layer during training to maximize information content. Instead of computing the representativeness as a cosine similarity between the descriptors of images, a content distance was introduced:

$$d_{cont}(I_i, I_j) = \frac{1}{N}\sum^{N}(R^l(I_i) - R^l(I_j))^2 \qquad (6)$$

   where layer activation responses $R^l(I_i)$ at the abstraction layer were used.
2. Instead of first selecting uncertain samples and then choosing the most representative of those, a Borda-count based method was proposed: samples are ranked for each metric, and the next query sample is picked based on the best combined rank.

The proposed strategy was tested on an MR dataset of 36 patients diagnosed with rotator cuff tear, where the goal was to segment bones and groups of muscles. Being initially trained on 64 slices the Dice score reached the upper bound (that is a result of training on 100% of the data D_{pool}) by using 27% of the D_{pool} .

3. Material and methods

In order to empirically examine behaviour of different active learning strategies we have created a synthetic dataset (See section 3.1), which served as a toy test problem during the algorithm development. Later we evaluate the methods on a lung nodule segmentation task using the publicly available LIDC dataset (See section 3.2).

3.1. Synthetic data

There exist different publicly available datasets that are used for deep learning experiments. MNIST by LeCun and Cortes (2010) is a well-known dataset that is used for handwritten digit classification, ImageNet by Russakovsky et al. (2015) is widely used for image classification. However, when it comes to segmentation task there is no one particular dataset that can be used for validation of different techniques. Therefore, we propose to create a synthetic dataset that can be used for segmentation and active learning experiments.

For an individual image the idea was to simulate a lung, nodules and other anatomical structures of different shades. To make the task more complicated for the network, samples with structures of the same shape as nodules but located outside the lung were generated. Examples of images are given in Figure 2.

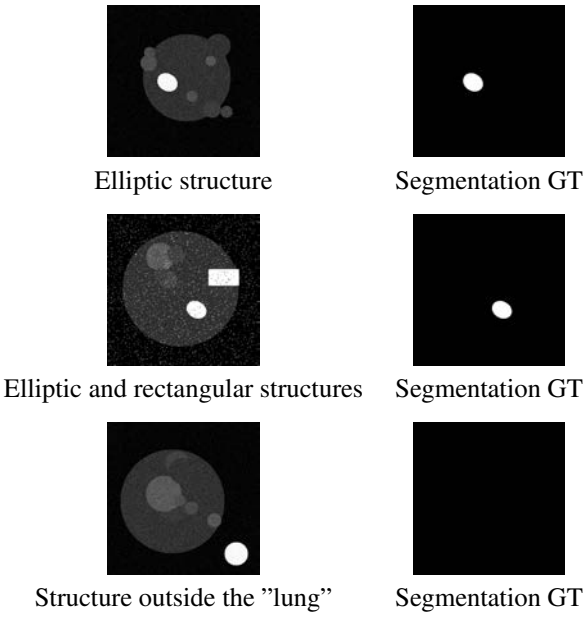


Figure 2: Examples of synthetic data

The distribution of types of generated images is given in Table 1.

Table 1: Types of synthetic data

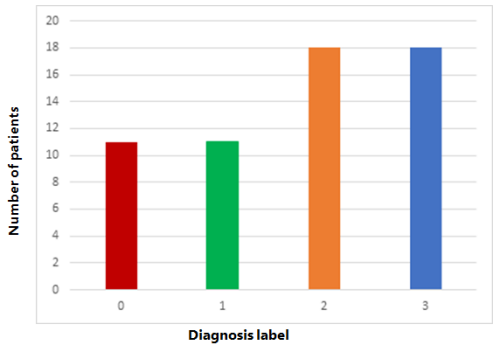
Type of structure	Location relative to the ”lung”	
	Inside	Outside
Round shaped	50	50
Elliptic shaped	50	50
Circles and ellipses with rectangles	50	50
Squares only	25	25

All of these 350 images were put into a training set. The task for the network is to correctly segment 150 round and elliptic structures inside the ”lung” and not to segment the same kind of structures if they are located outside. Bright squares and rectangles were added in order to add confusing elements that the network should learn as structures which must not be segmented. Additional 30 representative images were generated for the test set.

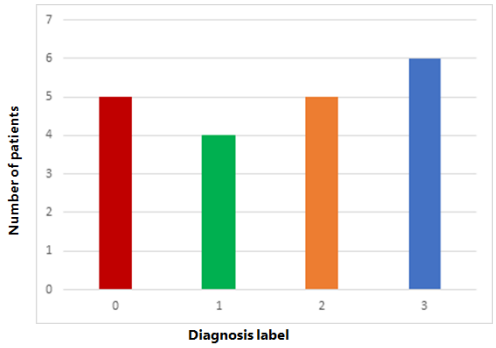
3.2. Medical data

In order to evaluate active learning strategies on computed tomography (CT) data we used scans from the Lung Image Database Consortium image collection (Armato III et al. (2011)). The database consists of 1018 cases including a clinical thoracic CT scan and an associated file with records of the results. The subset of the *LIDC-IDRI* with 483 patients formed a dataset for our

experiments. For every patient we have different number of slices. For each patient we extract all slices that have nodules (at least one per patient) and one random slice without any pathologies [2]. Additionally, we had information for 58 patients about their diagnosis, and we assigned part of those patients into a test set so we can analyze if active learning shows different results depending on the patient condition. Diagnosis distribution for all 58 patients and for a test set are given in Figure 3.



Diagnosis distribution for all available patients



Diagnosis distribution for the test set

Figure 3: Patients diagnosis. 0 - unknown, 1 - benign or non-malignant, 2 - malignant metastatic, 3 - malignant, primary lung cancer.

Details about the sets are given in Table 2.

Table 2:

Set	Number of patients	Number of slices	Number of nodules
Train	262	3 000	2 973
Validation	101	1 000	991
Test	20	246	239

3.3. Data pre-processing

For the synthetic data, pre-processing steps were simple: we resize images to 224x224 and normalize the values between 0 and 1.

When it comes to medical images at the pre-processing step we would like to enhance the contrast of nodules and clamp intensities of other anatomical structures. For this purpose HU values of the area inside the provided masks were extracted, and histograms of minimum and maximum values for all available data were created (Figure 4).

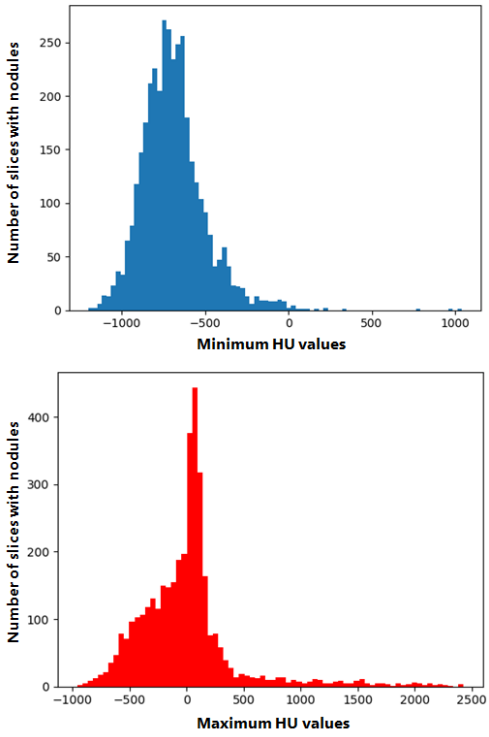


Figure 4: Histograms of maximum and minimum values of nodule areas

The windowing range was defined as a mean value +/− one standard deviation and was set as $[-882; 431]$. After windowing all values were normalized and resized to 320x320. Examples of images before and after preprocessing are given in Figure 5.

3.4. Network architecture

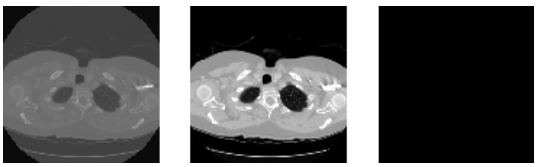
For experiments with both synthetic and real data U-Net like architecture was used. One of the biggest advantages of the network is that it proved to work well even with relatively small number of images.

In order to perform segmentation on synthetic data we do not need that much filter maps, and we propose the simplified shallow version of the original U-Net by Ronneberger et al. (2015) at Figure 6. The task for the synthetic data is not complex, and we expect the network to learn mostly intensity-based features, corners, lines.

In the work of Iglovikov and Shvets (2018) it has been demonstrated that U-Net performance could be

[2] We did not used all the available image slices for each patient as the number of the pixels that are non-nodule are an in overwhelming majority. By considering mostly nodule slices we force the model to be more sensitive than specific.

Slice of a patient LIDC-IDRI-0175 with a nodule



Slice of a patient LIDC-IDRI-0175 without nodules

Figure 5: From left to right: CT scan before preprocessing, CT scan after preprocessing, ground truth mask

improved by using pre-trained weights. Usually networks use weights trained on ImageNet as an initialization for other tasks. The modified version of U-Net has a VGG11 neural network (VGGU-net) without fully connected layers as an encoder. The VGG part contains seven convolutional layers with 3x3 kernels, each followed by a ReLu activation function, five max-pooling layers with a 2x2 pooling window. The last convolutional layer has 512 channels and serves as a bottleneck central part. For the decoding part, transposed convolutional layers that double the size of a feature map were used while reducing the number of channels by half. The output of a transposed convolution is then concatenated with the output of the corresponding encoder part. At the final layer a 1x1 convolution is used followed by a sigmoid activation function so we can get a probability of a pixel to be one of the two classes. The architecture of the network is given in Figure 7.

Since there are five max-pooling operations the input data should be divisible by 2^5, which is why we resize the images to 320x320. In the original paper the network was tested on RGB images (an input image consequently had 3 channels). We keep the same shape of input data, providing not only the slice containing nodules, but also slices before and after that one to give additional context.

3.5. Network parameters for training

It is important to apply a correct loss function for the most effective training. In case of large class imbalance focal loss proposed by Lin et al. (2017) (a modification of a standard cross entropy loss that down-weights the loss assigned to well-classified examples) maintains manageable balance between foreground and background. The focal loss is defined as:

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma} log(p_t) \tag{7}$$

where

$$p_t = \begin{cases} p, if\ y = 1 \\ 1-p, otherwise \end{cases} \tag{8}$$

with p being the model's estimated probability; $\alpha_t \in [0,1]$ is a weighting factor defined analogously to the p_t; γ is a tunable focusing parameter. We set $\alpha = 0.75$ giving more weight to a positive class and $\gamma = 2$ as suggested in the original paper for the experiments.
For both the shallow U-net and the VGGU-net the Adam optimizer with a learning rate of 10^{-3}, a batch size of 16 were used.

3.6. Metrics

For the evaluation on synthetic data we use a Dice score - a popular evaluation metric to quantify the performance of image segmentation.
$Dice = \frac{2|X \cap Y|}{|X|+|Y|}$, where X is the obtained segmentation and Y is the ground truth.

When we deal with very imbalanced classes[3] Precision-Recall might be a useful metric for evaluation. Precision is defined as $P = \frac{T_p}{T_p+F_p}$ and recall is defined as $R = \frac{T_p}{T_p+F_n}$, where T_p, F_p and F_n are true positives, false positives and false negatives respectively. In the case of an ideal classifier, it would return all results labeled correctly. The relationship between recall and precision can be observed in the area under the curve for different probability thresholds. Average precision (AP) summarizes such a precision-recall curve and is defined as:

$$AP = \sum_n (R_n - R_{n-1}) P_n \tag{9}$$

where P_n and R_n are the precision and recall at the n^{th} threshold. Further we refer to this metric as the PR score.

3.7. Active learning experiments with synthetic data

Based on the literature review, we would like to implement and compare active learning strategies described in the related work section. Despite Learning Active Learning being a promising technique, we did not include it in the comparison as it was implemented to work with random forest classifiers and has not been tested with neural networks.

The following active learning strategies were implemented:

1. Random sampling.

[3] there are way fewer nodule pixels than normal or background pixels in patients CT scans.

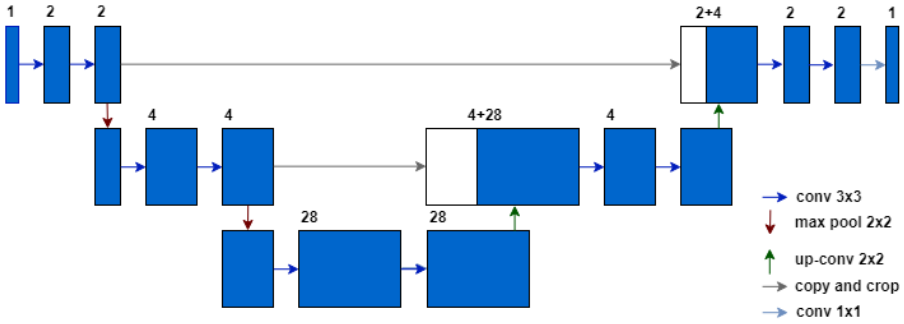


Figure 6: Shallow U-Net

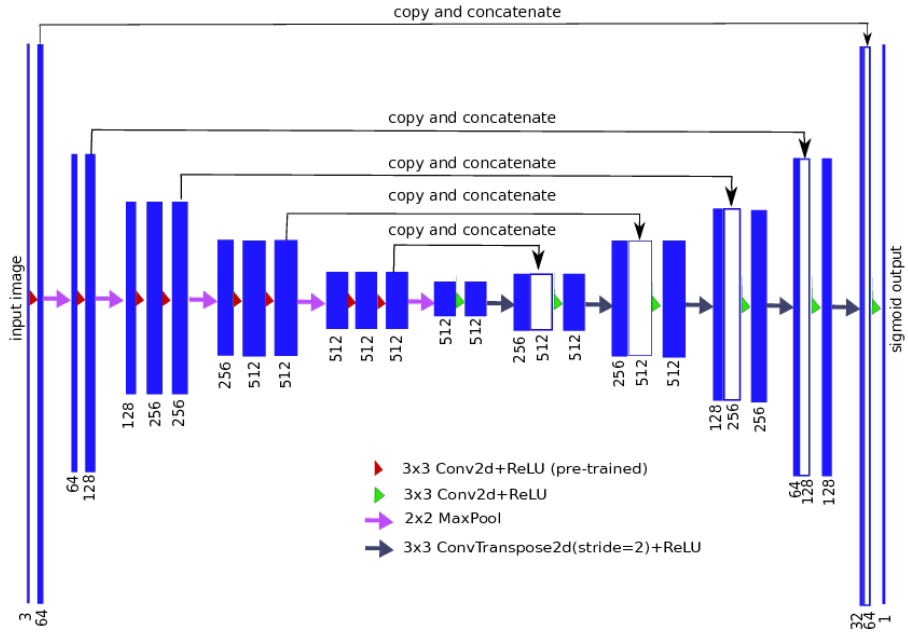


Figure 7: U-Net with VGG11 Encoder

2. Uncertainty sampling.
   One of the successful and simple strategies that focuses on selecting samples which the current network is the least certain about. The level of uncertainty is estimated based on using the effect of dropout. We propose two ways of computing the uncertainty level:
   - uncertainty-sum: sums up variances of predictions for each pixel
   - uncertainty-topmean: takes mean of 10% top most varying pixels
3. Uncertainty sampling + representativeness.
   With this strategy we estimate the similarity of the most uncertain samples with the unlabeled pool of data using cosine similarity.
4. Uncertainty sampling + pseudo-labeling.
   Since at the beginning when trained on very small amounts of data the network cannot produce correct segmentation masks, we add a condition that the average Dice score (measured on test data for active learning evaluation after each iteration) should be more than 0.8.
5. Uncertainty sampling + representative pseudo-labeling.
   Here we add an additional condition for assigning pseudo-labels. We compute cosine similarity between the most certain samples and samples in a training set, and choose those certain samples which are very close to what the network has seen before.

With two options for uncertainty estimation and two implementations of assigning pseudo-labels we have 9 strategies for active learning.
Each experiment has a pipeline as described below:

1. Randomly select 10 images as the training set and 35 images for validation.
2. Train a network with 10 images (for 1000 epochs using early stopping).

3. Based on a strategy choose two most uncertain images (and two most certain if the conditions are met for pseudo-labeling), add them into a training pool.
4. Continue training a network with images added using the same parameters as for initial training.
5. If mean Dice score for test images < 0.98 repeat steps 3 & 4 until the performance reaches or beat the upper bound.

For the first experiment we would like to prove the concept of applying a dropout layer at the end of the encoder part as an efficient way to estimate the level of uncertainty. We assume that if the network is trained only with images of a certain type, then during active learning iterations it should be very uncertain about images it has not seen before and therefore the most uncertain cases should be picked up for labeling and added into the labeled pool.
The sanity check experiment includes the next steps:

- Form a training data set with images that have rounded and elliptic shapes of "nodules" inside the "lung", as well as images with additional bright rectangular elements at random locations
- Train the shallow U-Net with all available data for receiving an upper bound, get mean Dice score for every type of image
- For the active learning experiment add a small amount of data with "nodules" of a circular shape to the initial training set
- At every active learning iteration choose two most uncertain images from the unlabeled pool of data, add them into the training set and continue training the network
- Analyze type of chosen data as the most uncertain, and track the number of samples in a training set when it reaches the upper bound

For that experiment we enlarged the data set to clearly see that it is possible to reach the highest performance with less data. In total 420 images were generated with equal distribution in types, then 281 images (67%) were put into the training and the rest into the validation set. 30 representative images were put into the test set that help track changes in the Dice score after every active learning iteration.

3.8. Active learning experiments with medical data

In this section we are providing implementation details of active learning strategies for the real data since they are slightly different.
For the upper bound, the network was trained several times with all 262 patients, and demonstrated an average PR score in a range of $0.69 - 0.72$.
For the AL experiments initially the VGGUnet was trained on ten randomly chosen patients, and the remaining 252 patients were put into the unlabeled pool of data.
Four main active learning strategies were implemented.

1. Random sampling.
   At every active learning iteration four random patients are chosen from the unlabeled pool and put into a training set.
2. Uncertainty sampling.
   For every image pixel I_x we compute the variance of T different predictions I_y on the same pixel. For every patient there is a different number of samples, and in our implementation we propose the following algorithm:
   - For every sample from a patient compute predictions three times
   - Compute variance for every pixel and compute mean of all variances to obtain a numerical score of the uncertainty level per sample
   - Take mean value of all scores in order to get the level of uncertainty for a patient

   Once the levels of uncertainty have been computed for all patients from an unlabeled pool, we take four the most uncertain patients and add them to the training set.
3. Uncertainty sampling + Representativeness.
   In order to define the four most representative patients we follow the next steps:
   - After estimating the level of uncertainty take 8 most uncertain patients as potential candidates.
   - Estimate how similar every most uncertain patient is with the rest of the patients from the unlabelled pool: for every sample of the most uncertain patient compute cosine similarity with every sample of an unlabelled patient (thus we estimate how similar two patients are)
   - Taking the mean value of all similarities per patient we measure the level of representativeness of a given patient to the unlabeled pool

   Add the four most uncertain and representative patients to the training set.
4. Uncertainty + Pseudo-labeling.
   In order to make sure that automatically computed masks are close enough to the ground truth we add two conditions that must be satisfied before assigning pseudo-labels:
   - The average PR score must be greater than 0.65
   - Mean uncertainty level per patient must be smaller than 0.01.
     That level was empirically received by computing the uncertainty level of patients from a test set for which we observed high PR score with the network fully trained on the whole available training set (3 000 samples).

In order to estimate the network performance at each active learning iteration, we compute the average precision-recall score using the test set.

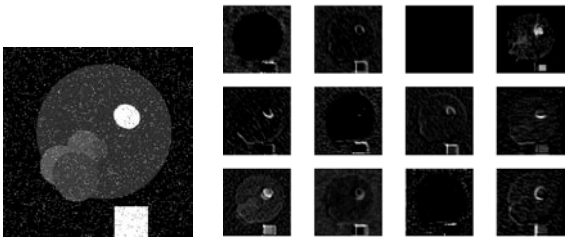


Figure 8: Example of extracted feature maps for synthetic data

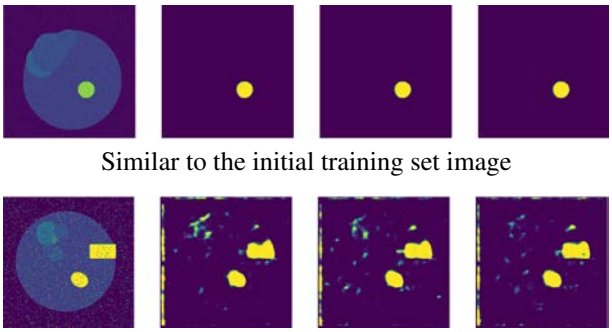


Figure 9: Predictions for similar and new images: left image - original sample, three images next to it - predictions after 1 AL iteration with enabled dropout layer

4. Results

4.1. Proof of concept on synthetic data

For the first experiment we had to train the shallow U-net with all labeled images to receive the upper bound. The network managed to successfully extract expected features of the image such as shape and intensities. In Figure 8 you can see an example of 12 out of 28 feature maps extracted at the bottom layer of a shallow U-Net.

The high Dice score was obtained for every type of images (Table 3).

Table 3:

Type of images	Mean Dice score
Round shaped	0.981
Elliptic shaped	0.979
Circles and ellipsis with rectangulars	0.977
Mean Dice	**0.979**

Then for the active learning experiment initially the network was trained with 50 images that contain only ”nodules” of a round shape. Our assumption was that in this case even with an enabled dropout layer the network should still be confident about predictions for similar images and should demonstrate variant predictions for images with squares and added noise. After the first AL iteration we got three predictions with an enabled dropout layer and results for an image similar to the training set and an unseen image are shown in Figure 9.

After running an active learning experiment based on estimation only of the level of uncertainty we noticed gradual improvements in the Dice score for every type of images as shown in Figure 10.

The graph at Figure 11 demonstrates the progress of a mean Dice score showing the most uncertain images that were added into the training set at iterations number 1, 10, 20, 30, 40 and 50.

4.2. Active learning experiments with synthetic data

Every active learning strategy was tested three times with different samples being put into the initial training set, but for every experiment the initial set was the same. You can find results for every experiment in Figure 14 and the images that were randomly put to the initial training sets in Figure 13. The mean Dice score per strategy is summarized in Figure 12.

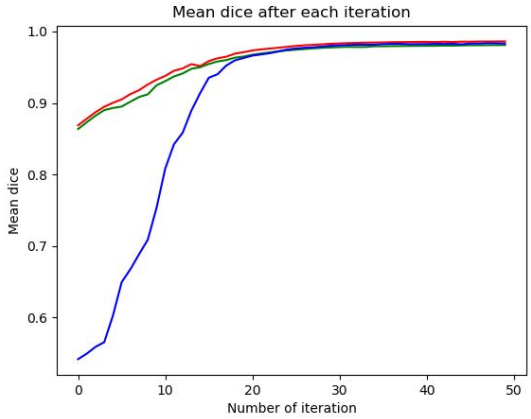


Figure 10: Mean Dice score per type of image at every AL iteration: red - round nodules, green - elliptic nodules, blue - images with noise and rectangular structures

4.3. Active learning experiments with medical data

Results for the first active learning experiments with LIDC-IDRD data set are given in Figure 15. The graph was obtained only for one random split into initial training set and unlabeled pool of data. After few alterations in the network parameters and AL algorithms that will be explained in the next section, the new robust results were obtained (Figure 16). The results of the winning algorithm at patient level is given in Figure 18. We also investigated what kind of samples had the highest impact on the performance. Few CT scans of such helpful patients with their masks are shown in Figure 17. Figures 19 - 22 demonstrate segmentations of the winning strategy for patients with different diagnosis, and Figures 23 - 24 show examples of assigning pseudo-labels at early iterations.

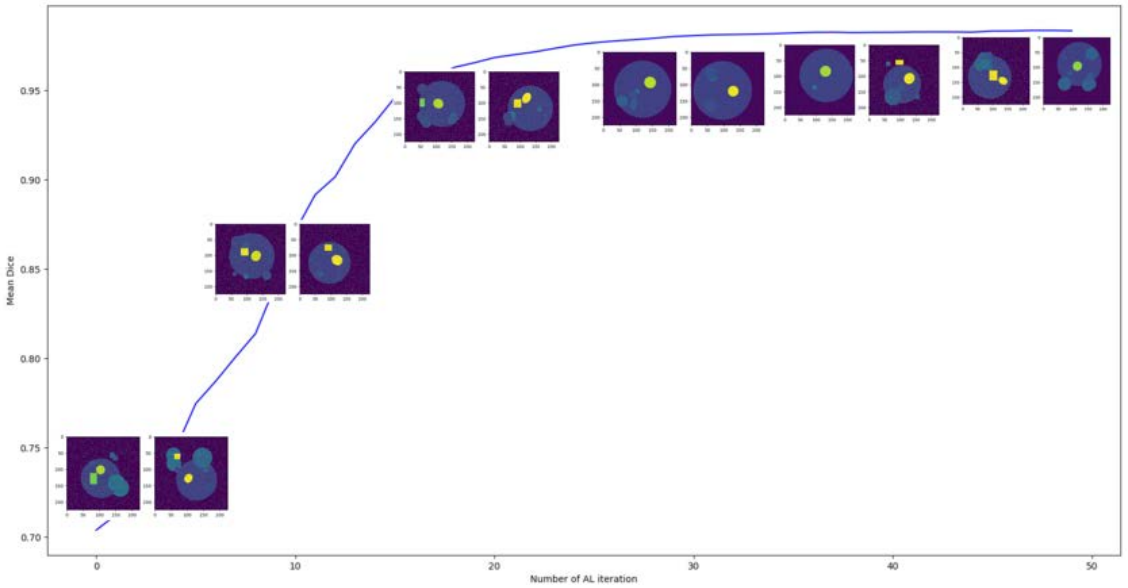


Figure 11: Mean Dice score at every AL iteration with chosen most uncertain images

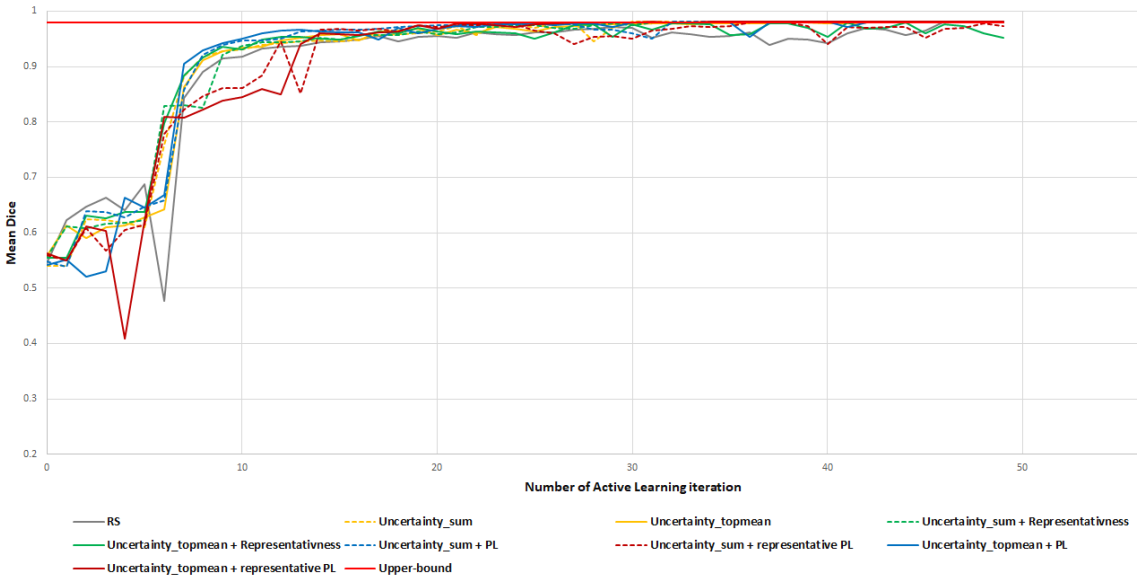


Figure 12: Mean Dice score per strategy

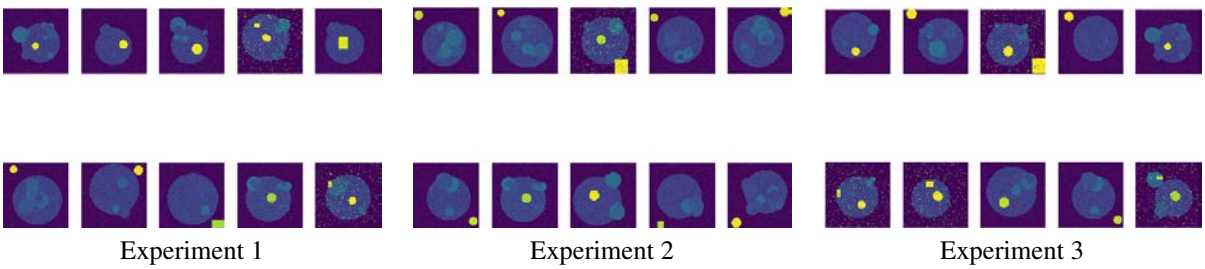
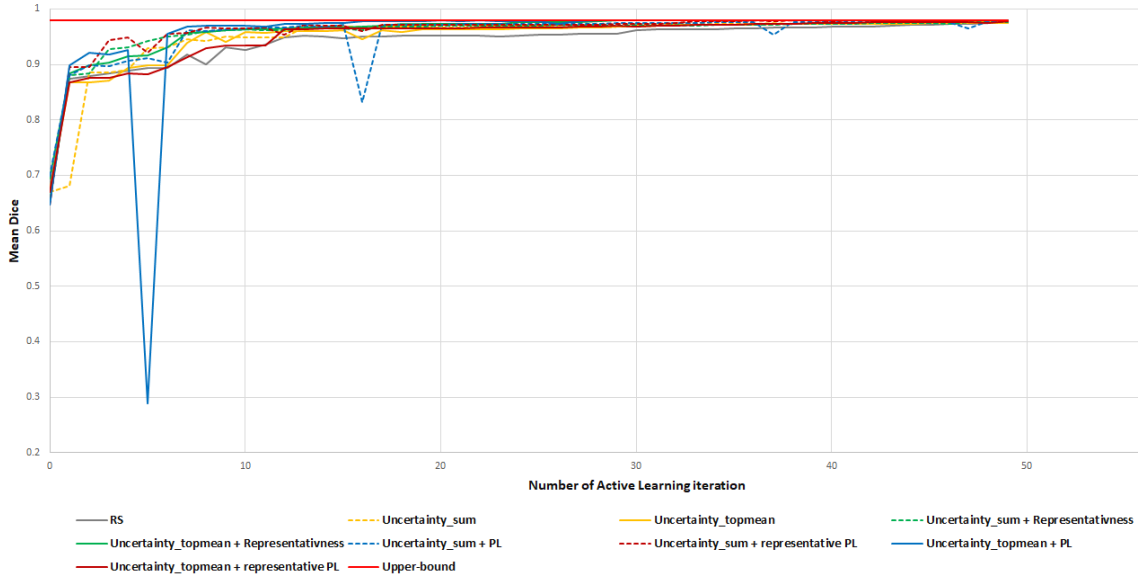
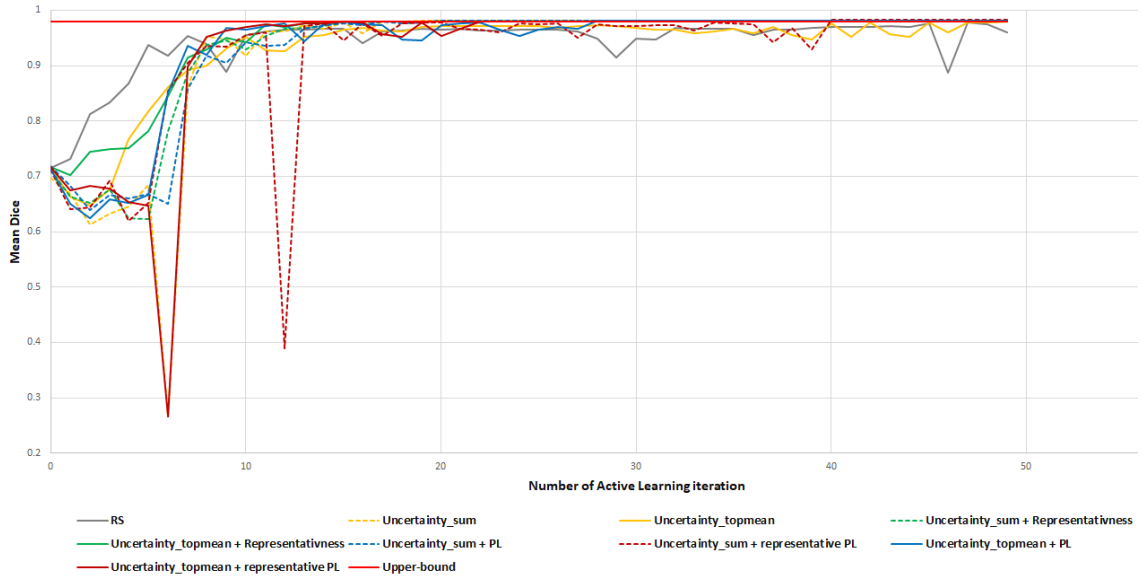


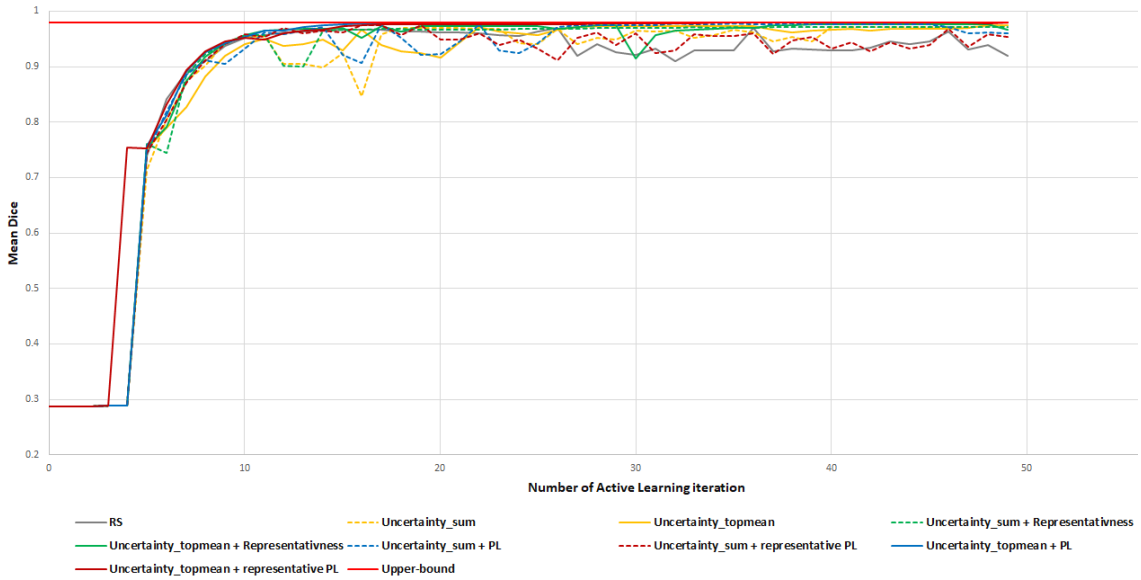
Figure 13: Initial training sets



Experiment 1



Experiment 2



Experiment 3

Figure 14: Active Learning results for synthetic data

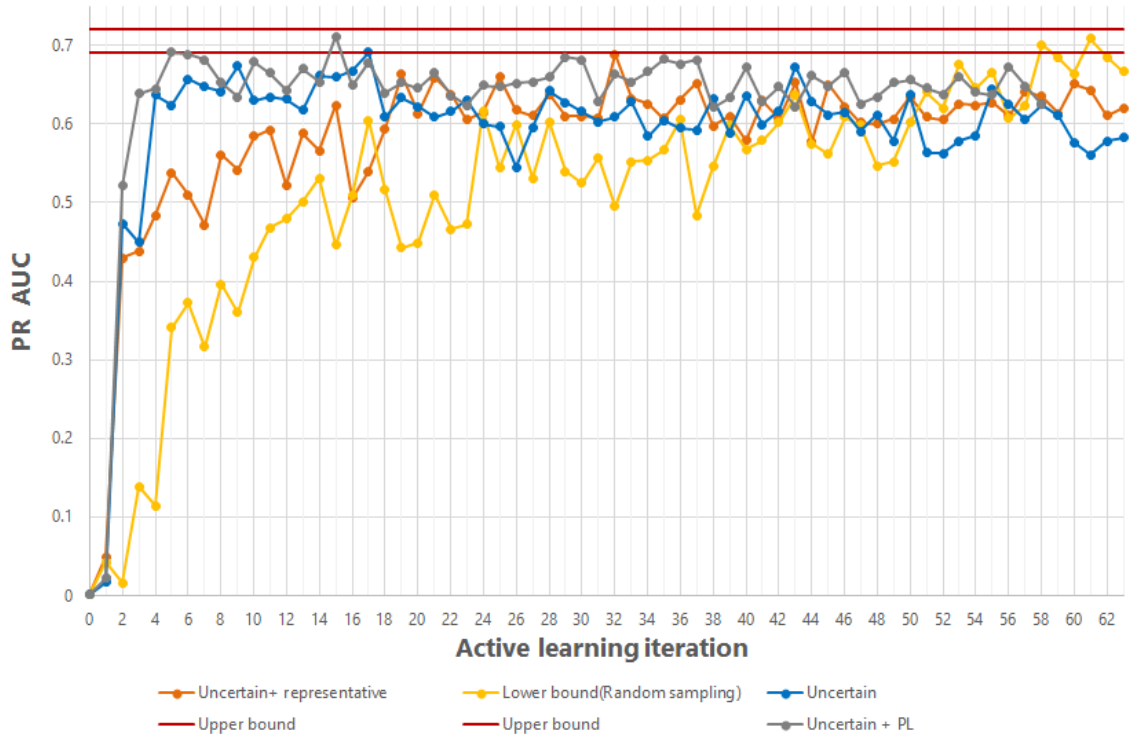


Figure 15: First results. Mean PR score for every strategy per AL iteration

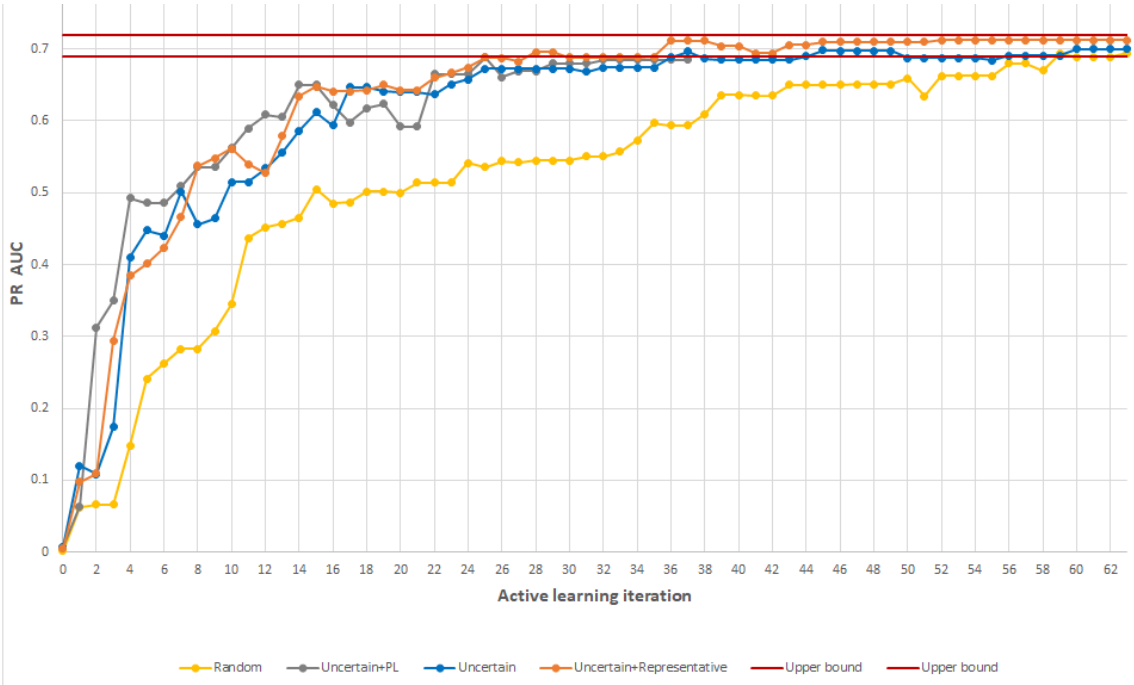
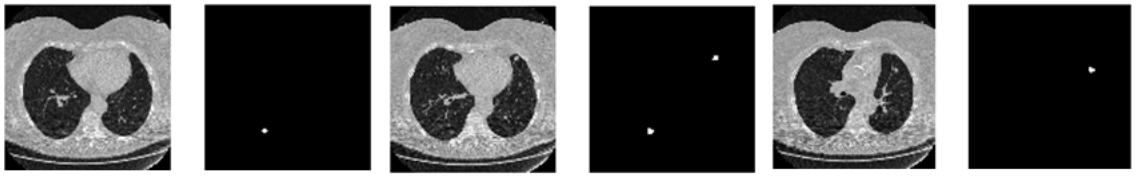
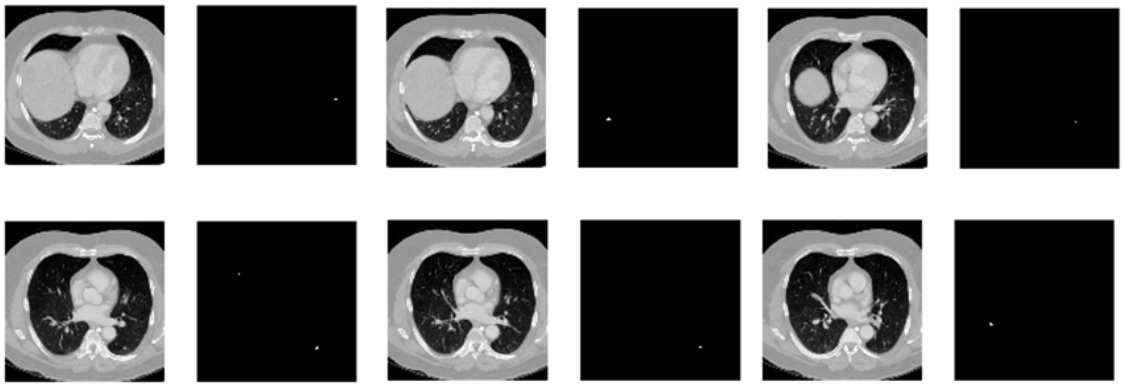


Figure 16: Final robust results. Mean PR score for every strategy per AL iteration



Slices with ground truth masks of Patient LIDC-IDRI-0476



Slices with ground truth masks of Patient LIDC-IDRI-0398

Figure 17: Example of patients which significantly contributed to the boost of performance

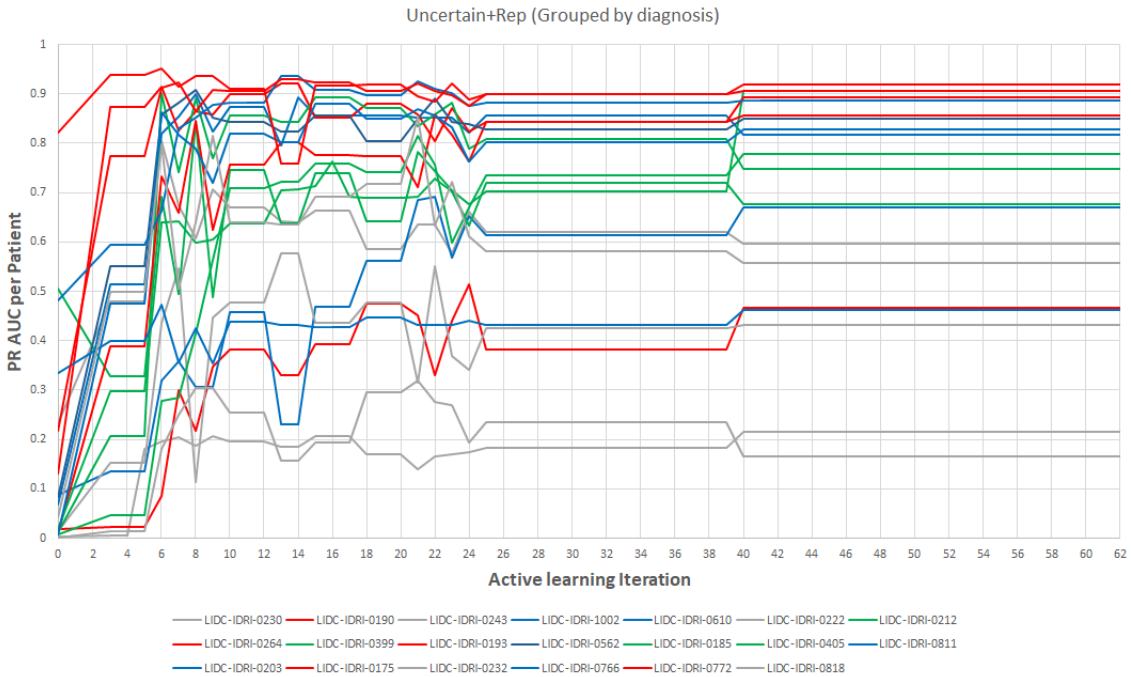


Figure 18: Mean PR score for Uncertainty+Representativeness at patient level. 3(red) - malignant metastatic; 2(blue) - malignant, primary lung cancer; 1(green) - benign or non-malignant; 0(grey) - unknown

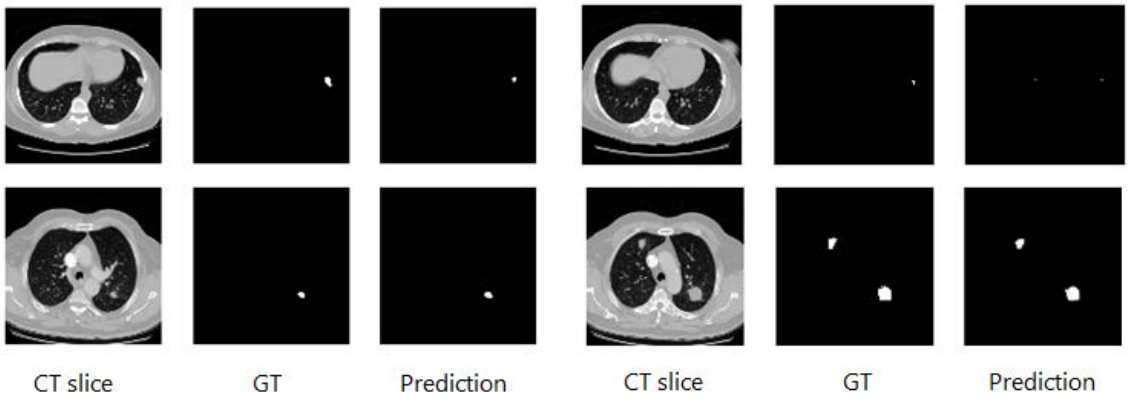


Figure 19: Segmentation results for malignant metastatic diagnosis. Poor result - on top, good - below

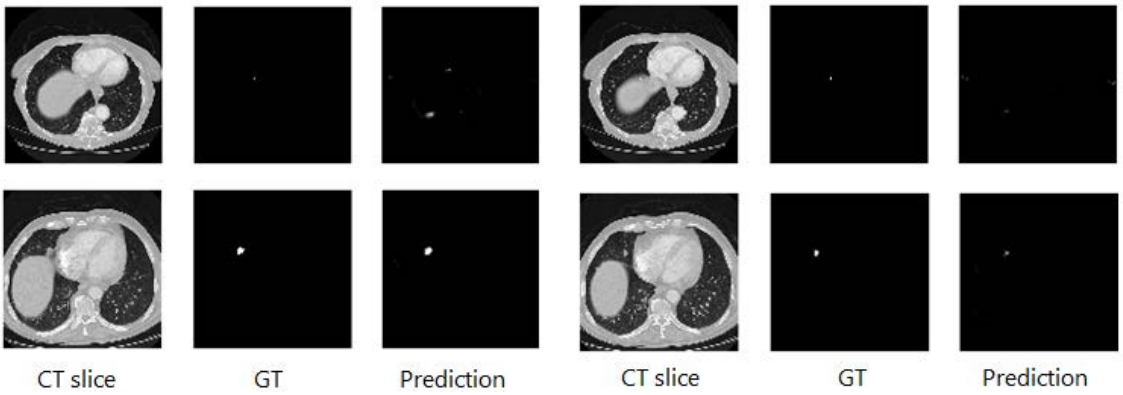


Figure 20: Segmentation results for lung cancer. Poor result - on top, good - below

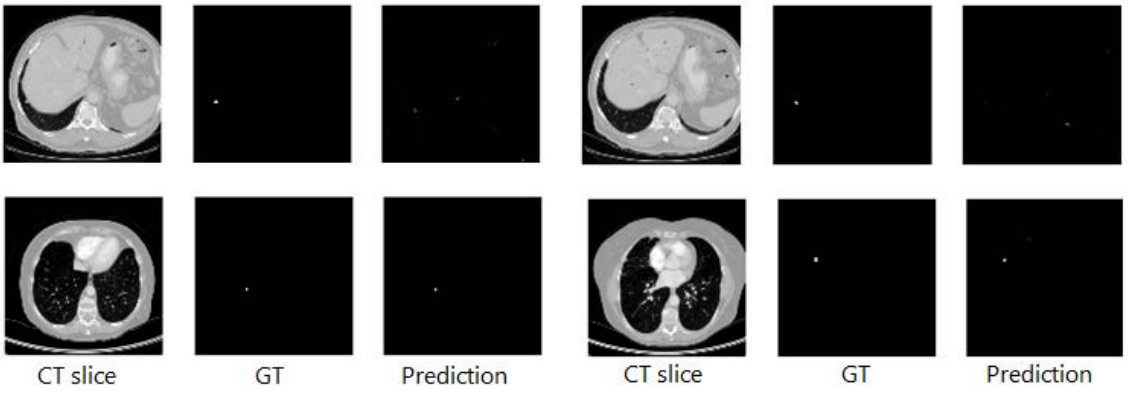


Figure 21: Segmentation results for benign. Poor result - on top, good - below

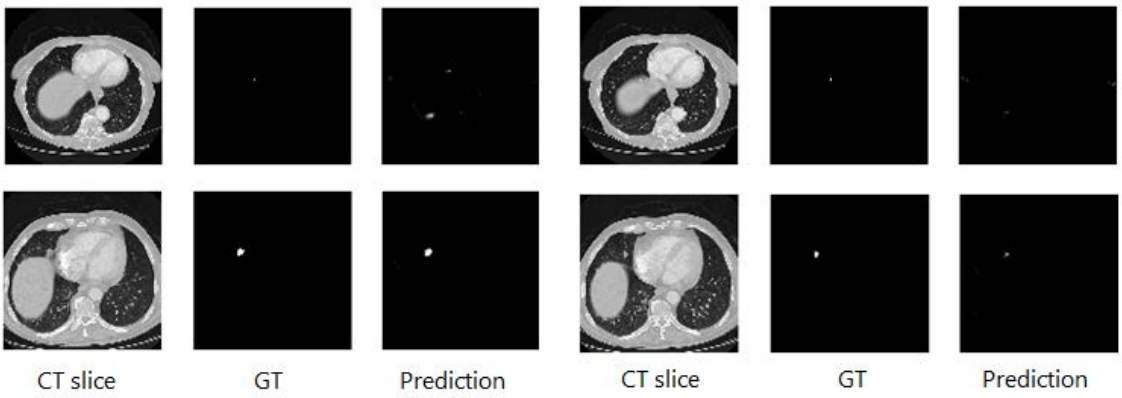


Figure 22: Segmentation results for unknown diagnosis. Poor result - on top, good - below

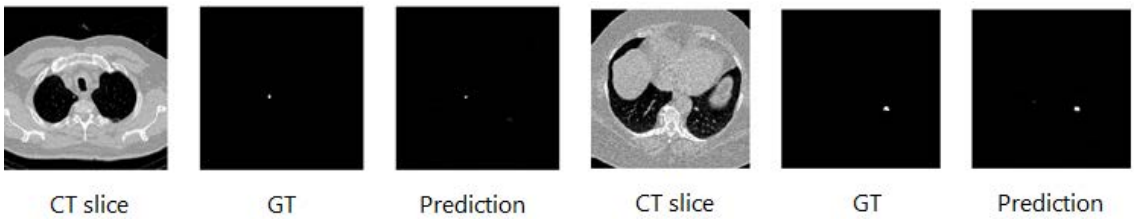


Figure 23: Most certain patients with correct pseudo-labels

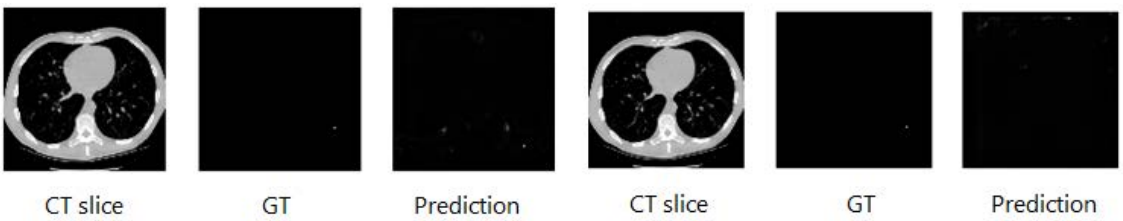


Figure 24: Most certain patients with incorrectly assigned masks

5. Discussion

5.1. Proof of concept on synthetic data

The first experiment with synthetic data showed promising results and proved the concept of applying the Dropout layer for the uncertainty estimation. Having been trained with one type of image (”nodules” of a round shape) the network successfully segmented a ”nodule” in the image that was similar to the initial training set. After applying the dropout layer, U-net still was confident and produced masks (Figure 9) which visually look very similar with minor differences at pixel level. For the previously unseen sample we may observe variations of predictions around borders and for structures inside a ”lung”.
Looking at Figure 11 we may conclude that estimating uncertainty level is a smart strategy, at first iteration it was choosing noisy images with rectangular structures close to the ”nodules”, and the mean Dice score for that type of image (blue line in Figure 10) was gradually increasing after adding two new samples. We can observe a great improvement after 20 iterations for all three types of images. The upper bound was reached after 26 iterations (with 50 + 26*2 = 102 images that is 36% of the available training data) and even outperformed the results, reaching a mean Dice score of 0.983 after 45 iterations (with 50 + 45*2 = 140 images that is half of the training data).

5.2. Active learning experiments with synthetic data

Main experiments demonstrated different results depending on what types of images we have in the initial small training set. In experiment 1 we have all types of images from the beginning, the initial set looks quite representative showing high starting Dice score. During the first iterations ”Uncertainty-topmean+PL (pseudo-labeling)” shows the highest Dice score, however at the fifth iteration we observe a great dip caused by assigning pseudo-labels which were not correct, meaning that only the condition for activation of pseudo-labeling was not enough and the network still was not trained well. Representative pseudo-labeling on the contrary managed to correctly produce masks for samples that look similar to the training set, and the overall performance was increasing for the next iterations. Other techniques showed small variations at the beginning in terms of performance, but they all managed to converge at around the 45th iteration.
For the second experiment negative samples (without structures that should be segmented, an imitation of images without nodules) prevailed in the initial training set. At first smart strategies tended to pick up samples with rectangular structures only inside the ”lung” as the most uncertain, they look similar to positive examples, but the shape is different and they are still belong no negative group of pictures. Without adding positive examples we obviously could not get improvement in segmentation. But once negative samples were learned by the network, smart strategies started to choose desired images with ”nodules” inside the ”lung” and rectangular anatomies that led to noticeable improvements after the 8th iteration for every strategy. Both pseudo-labeling techniques showed two great dips at the beginning, however the decrease in the mean Dice score was not caused by wrongly assigned labels: the solid red line which is related to representative PL at this iteration does not include most certain samples with their automatically produced masks and should behave as purely uncertainty-based algorithms, and when we checked samples that were added at iteration 12 for the dotted red line, then we could see negative samples as most certain with masks being correctly produced (without any pixels assigned to a class 1 as for nodules). We assume that such instability could be caused by a high learning rate of the network that stays unchanged while we are progressing with a training, by the time we noticed great dips in the training set a lot of negative samples were accumulated which probably made the network produce false negative results. The rest of the techniques show expected behaviour leaving random sampling a little beyond, all strategies managed to converge.
The worst starting point was for experiment 3: there were no negative samples with rectangular bright structures inside the ”lung”, and probably the network mistakenly segmented them as true samples. For all strategies but RS, images with round and elliptic structures inside the ”lung” were chosen as the most uncertain. After adding enough positive samples we may observe a big improvement for every algorithm after the 10th iteration, even RS by luck chose missing from the initial set types of images that led to increasing Dice score. Not all of the algorithms, however, could reach the upper bound, and for many of them there were fluctuations around a Dice score of 0.95, that again could be a result of a high learning rate.
Results of all experiments were put together as mean Dice score per algorithm in Figure 12, we were wondering if we may see a clear winning strategy. There is no noticeable difference in the way we estimate the level of uncertainty, so we stick with one way of defining uncertainty for the medical data. However, mean results do not help us in defining the best algorithm that reached the upper bound fastest, all strategies require a few iterations at the beginning to collect valuable samples before the boost in performance. From iteration 20 to 25 all methods managed to get close to the upper bound with 50-60 images in a training set which is only 15-20% of the whole available data for training.

5.3. Active learning experiments with medical data

The first results we obtained for real data look quiet noisy. Uncertainty and uncertainty+PL behave in a similar way in the beginning as expected because pseudo-labeling was not activated before reaching the threshold of average PR score. At the 15th iteration uncertainty+PL hits the upper bound first. It is interesting to notice, that for those patients which were defined as most certain, automatically produced masks (Figure 24) were not always correct. However, adding few samples with incorrect masks did not reduce the overall performance of the network. They may serve as regularization preventing the network from over-fitting.

Potentially adding not just uncertain but very representative cases should lead to greater improvements, and we would expect uncertain+representative to outperform pure uncertain methods which is not the case in our experiment. That fact allowed us to conclude that the way the representativeness was calculated was not the best choice. It was also computationally costly since every patient's slice was represented as a vector of 51200 elements. The new proposal was to apply PCA in order to project the huge feature vector to a lower dimensional space. We applied kernel principal component analysis for non-linear dimensionality reduction with a cosine kernel and kept 10 main components.

Another concern was that only RS managed to converge in the end, but smart strategies after reaching the upper bound at early iterations, all started decreasing gradually with adding more data. The problem might be caused by over-fitting since in that experiment we kept the learning rate fixed in the same way it was done with synthetic data. Additionally, we noticed that only model's weights were saved after each iteration, but it is important to save the optimizer's state as well, so for the new active learning iteration the loss starts from its last position. We tried to use cycling learning rate proposed by Smith (2017) instead of a fixed one. The idea is that learning rate changes between a lower bound and upper bound. Despite the common rule that LR should decrease as training progresses, it might be reasonable to give the model higher LR periodically so that it helps to escape local minimas if it ever enters into one.

The pseudo-labeling technique showed great results in the first experiment, however at the beginning it behaved like the uncertainty algorithm and we cannot really say if assigning masks automatically led to the boost in performance. In order to investigate the impact of pseudo-labels we lowered the threshold for enabling pseudo labeling and let the algorithm start assigning masks almost at the beginning (the new threshold was set at 0.1, for variance it stayed the same 0.01).

We managed to run experiments for every strategy three times and Figure16 shows mean results. We can clearly see that random sampling performs worse than smart techniques and serves as a lower bound. Combination of uncertainty estimation with representativeness slightly outperforms pure uncertainty and first reached the upper bound at iteration 25. Pseudo-labeling (PL) takes the leading position during the first iterations which is caused by adding more patients to the training data than for other algorithms (in best case if all conditions are met, then additionally to the four most uncertain patients it chooses the four most certain patients with their predictions to be added to the training pool), despite lowering the first threshold and training the network only with few patients, it managed to produce good masks for most certain patients (Figure 23). PL reaches the upper bound at the same iteration as Uncertain+Representativeness (U+R), however during the next iterations performance is declining slightly, and such behaviour is predictable since we do not take out those samples that were automatically annotated and we keep their predictions (not always correct ones) in the training data.

The way an active learning strategy should be evaluated depends on how much data were required when the performance of the model managed to reach an upper bound. U+R and PL both reached upper bound at iteration 25, however the first algorithm used less data at the same iteration, and showed further improvements in performance with more data added at iteration 36. The best result was achieved by U+R with 58% of the available labeled data for the training. A summary is given in Table 4.

For every strategy we noticed big improvements in average PR score at early active learning iterations, we checked for every strategy what patients were added and found some overlapping patients for different AL algorithms that led to increasing of PR score. For example, patient LIDC-IDRI-0476 (Figure17) with 23 slices has nodules of a very big size, and patient LIDC-IDRI-0398 with 28 slices has nodules of less diameter but their number is high, they are located at different places, so that patient gave a lot of useful information. It was a general trend for all strategies that adding patients with more slices and bigger nodules [4] resulted in better segmentation.

Analyzing results for the U+R algorithm at a patient level, we may conclude that the best segmentation was achieved for patients with malignant metastatic diagnosis, but even for the patient that stands out on a graph with a low average PR score with the same diagnosis (0.47) we may see that nodules were detected but undersegmented with a small false positive area (in Figure 19 we can compare good and poor segmentation re-

[4] Added samples with large nodules helped more at earlier iterations of the active learning strategy and had less impact later on. This resembles a curriculum learning strategy, learning to segment easy cases first and gradually adding more challenging samples.

Table 4:

AL algorithm	Iteration	Initial # of annotated patients	# of patients annotated by a specialist	# of automatically annotated patients	Total # of patients	% of used data from the whole available labeled data for training
U+R	25	10	25*4 = 100	n/a	110	42%
U+R	36	10	36*4 = 144	n/a	154	58%
U+PL	25	10	25*4 = 100	15*4 = 60	170	65%

sults for two patients with malignant metastatic diagnosis). Cancer (Figure 20) was also detected well since nodules are mostly large in size. Patients with benign tumors showed high stable PR score, despite the small nodule size, the network managed to properly segment tiny nodules failing only one case for few CT scans (Figure 21). Unknown diagnosis means that a biopsy was not carried out and we do not know for sure the type of nodules. That is the group of patients with the highest variability in PR score with many false positive results. Since AL algorithms were choosing patients with large nodules to be put into the training set at the beginning, the network learned faster how to properly segment malignant cases.

6. Conclusions

In this study, we compared results of different active learning techniques. Based on existing works in this field we implemented the most popular algorithms, however, some changes in the implementations were required to adapt them to the problem of lung nodule segmentation with VGGU-net. We proved the main concept of AL with synthetic data and showed on real cases that with less data it is possible to reach high network performance.

For the future work it would be interesting to investigate more about a pseudo-labeling technique. Better segmentation results could be achieved if we took automatically annotated samples out of the training set and clear their predictions after every AL iteration, then with the progress of a network performance we could get more accurate predictions for most certain patients that potentially should lead to a boost of performance at later iterations. Moreover, we restricted a maximum number of the most certain patients to be added to four, but we might get more samples with the mean variance less than a threshold. Since a pseudo-labeling does not require human labour, we may add all patients the network is certain about and obtain boost in performance at the beginning.

The training itself may be more efficient by learning from unlabeled data. In the work of Jean et al. (2018) additional unsupervised loss term is proposed that minimizes the posterior variance at unlabeled data points.

There is still a room for improvement in terms of segmentation results, but we managed to reach satisfying results with way less data (the range varies from 42% to 65% depending on the chosen strategy) than it was done by using the whole available annotated set.

7. Acknowledgments

First of all, I would like to thank my supervisor Aneta Lisowska for her guidance, support and scientific advice. I would also like to express my gratitude to the whole AI team at Canon Medical Research Europe company in Edinburgh, United Kingdom for providing an excellent placement, warm working environment and top-level equipment that made it possible to perform all planned experiments.

I would also like to thank all the professors from University of Burgundy, University of Cassino and Southern Lazio and University of Girona for their interesting lectures, knowledge they shared and constructive feedback. To all my MAIA family, thank you for being such an important part during these 2-year program and wishing you best of luck in your future career.

References

Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al., 2011. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics 38, 915–931.

Gorriz, M., Carlier, A., Faure, E., Giro-i Nieto, X., 2017. Cost-effective active learning for melanoma segmentation. arXiv preprint arXiv:1711.09168 .

Iglovikov, V., Shvets, A., 2018. Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation. arXiv preprint arXiv:1801.05746 .

Jean, N., Xie, S.M., Ermon, S., 2018. Semi-supervised deep kernel learning: Regression with unlabeled data by minimizing predictive variance, in: Advances in Neural Information Processing Systems, pp. 5322–5333.

Konyushkova, K., Sznitman, R., Fua, P., 2017. Learning active learning from data , 4225–4235.

LeCun, Y., Cortes, C., 2010. MNIST handwritten digit database URL: `http://yann.lecun.com/exdb/mnist/`.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988.

Ozdemir, F., Peng, Z., Tanner, C., Fuernstahl, P., Goksel, O., 2018. Active learning for segmentation by optimizing content information for maximal entropy, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 183–191.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) 115, 211–252. doi:10.1007/s11263-015-0816-y.

Scheffer, T., Decomain, C., Wrobel, S., 2001. Active hidden markov models for information extraction, in: International Symposium on Intelligent Data Analysis, Springer. pp. 309–318.

Settles, B., 2009. Active learning literature survey. Technical Report. University of Wisconsin-Madison Department of Computer Sciences.

Shannon, C.E., 2001. A mathematical theory of communication, vol. 5 new york. NY: ACM SIGMOBILE Mobile Computing and Communications Review , 3–55.

Smith, L.N., 2017. Cyclical learning rates for training neural networks, in: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE. pp. 464–472.

Wang, K., Zhang, D., Li, Y., Zhang, R., Lin, L., 2016. Cost-effective active learning for deep image classification. IEEE Transactions on Circuits and Systems for Video Technology 27, 2591–2600.

Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z., 2017. Suggestive annotation: A deep active learning framework for biomedical image segmentation, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 399–407.