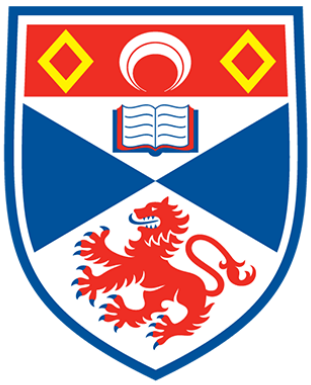
Segmentation of Breast Cancer on Contrast Enhanced Spectral Mammograms Using Affine Data Augmentation Techniques

Sean Alger



University of
St Andrews

Supervised by Dr David Harris-Birtill

BSc Computer Science Dissertation, University of St Andrews

March 2024

Abstract

The segmentation of medical images plays a crucial role in the detection of breast cancer in mammograms. Accurate segmentation facilitates early diagnosis, which allows personalised treatment plans to be given to the patient, thereby improving patient outcomes. However, the training of these segmentation deep learning models requires large datasets to ensure the best results. Unfortunately, many publicly available datasets do not contain enough images to train deep learning models to their full potential. This project aims to investigate the effects that different affine data augmentations have on the Dice Score of a U-NET model when utilising a new public dataset containing Contrast-Enhanced Spectral Mammography (CESM) images [1]. This dataset contains a total of 1003 CESM images and corresponding segmentation masks made by a trained radiologist. The investigations involve tuning model parameters on the CESM images and then exploring the effects that individual and combined data augmentations have on the performance of the model. The best-performing model included images that have been sheared vertically as well as translated in the x-direction. This model achieved a dice score of 56.6% on the test set, a 9% improvement from the baseline model, showcasing the importance that data augmentation has when using small datasets. This dissertation also serves as a baseline for the segmentation of images from this dataset, which future research can build upon.

Declaration

I declare that the material submitted for assessment is my own work except where credit is explicitly given to others by citation or acknowledgement. This work was performed during the current academic year except where otherwise stated. The main text of this project report is 15,138 words long, including project specification and plan. In submitting this project report to the University of St Andrews, I give permission for it to be made available for use in accordance with the regulations of the University Library. I also give permission for the title and abstract to be published and for copies of the report to be made and supplied at cost to any bona fide library or research worker, and to be made available on the World Wide Web. I retain the copyright in this work.

Acknowledgements

I would like to thank my supervisor, David Harris-Birtil, for his continuous guidance, encouragement, and expertise throughout the entirety of the dissertation process. David's advice and recommendations were indispensable and essential for the completion of the project. In addition to that, I would like to thank all the staff for the incredible teaching during my four years of study in the department of Computer Science. Finally, I would like to thank my family and friends for their support, they have truly helped me push through my final year of study.

Contents

1 Introduction **6**
- 1.1 Objectives . . . 8
  - 1.1.1 Primary . . . 8
  - 1.1.2 Secondary . . . 8

2 Literature Review **9**
- 2.1 Non-Deep Learning Techniques in CAD . . . 9
  - 2.1.1 Classical Segmentation Methods . . . 9
  - 2.1.2 Machine Learning Methods . . . 11
- 2.2 The use of Deep Learning CAD . . . 13
  - 2.2.1 Convolutional Neural Networks . . . 13
  - 2.2.2 U-NET Architecture . . . 15
  - 2.2.3 Deep learning in the Segmentation of breast cancer . . . 16
- 2.3 Publicly Available Datasets . . . 20
- 2.4 Transfer Learning and Image Size in Digital Mammograms . . . 21
- 2.5 Data Augmentation . . . 22
- 2.6 Direction . . . 27

3 Ethical Considerations **28**

4 Methodology **29**
- 4.1 Approach . . . 29
- 4.2 Dataset . . . 29
- 4.3 Data Augmentations . . . 31
  - 4.3.1 Flipping . . . 31
  - 4.3.2 Rotation . . . 31
  - 4.3.3 Shearing . . . 32
  - 4.3.4 Translation . . . 32
- 4.4 Pipeline . . . 33
- 4.5 Technologies Used . . . 34
- 4.6 Performance Metrics . . . 35
  - 4.6.1 Accuracy . . . 35
  - 4.6.2 Dice Coefficient . . . 35
  - 4.6.3 Precision . . . 36
  - 4.6.4 Recall . . . 36

5 Validation Set Investigations **37**
- 5.1 Baseline Model . . . 37
- 5.2 Hyperparameter Tuning . . . 41
- 5.3 Loss Function and Optimiser Function . . . 44
  - 5.3.1 Dice Loss . . . 45
  - 5.3.2 Dice Binary Cross Entropy Loss . . . 45

5.3.3 Tversky Loss . . . 45
5.3.4 Experimental Results . . . 46
5.4 Seed Search . . . 46
5.5 Individual Data Augmentation Techniques . . . 49
5.5.1 Flipping Images . . . 50
5.5.2 Rotation of Images . . . 50
5.5.3 Shearing of Images . . . 51
5.5.4 Translation of Images . . . 51
5.5.5 Overall Effect of Individual Data Augmentations . . . 52
5.6 Combined Data Augmentation Techniques . . . 52

6 Test Set Results **55**

7 Discussion **61**
7.1 Critical Appraisal . . . 61
7.2 Limitations . . . 63
7.3 Future Work . . . 63
7.4 Evaluation . . . 64

8 Conclusion **66**

9 Appendix **74**
9.1 Tables of full results . . . 74
9.1.1 Hyperparameter Tuning Full Results . . . 74
9.1.2 Combining Data Augmentations Full Results . . . 76
9.2 Ethics Approval Document . . . 77

1 Introduction

There is an imperative for accurate segmentation models in breast cancer detection, due to its prevalence among women today [2]. Statistics published in 2021 reveal that breast cancer alone accounts for almost one-third of all cancer diagnoses in women [3]. In fact, 2.5% of fatalities among women can be directly attributed to breast cancer [2].To mitigate the high incidence and mortality rates associated with breast cancer, regular and consistent screening is essential. In line with this, the NHS recommends that women aged 50 and above undergo mammography screenings every 3 years [4]. Supporting this, a 2005 study reveals that screening mammography reduces breast cancer mortality by 20% to 35% in women aged 50 to 69 years old [5].

In the 20 years that digital mammography (DM) has existed, it has widely been used and has been called the gold standard for breast cancer screening [6]. However, its performance is not perfect, especially in the scanning of particularly dense breast tissue, which may make it harder to differentiate suspicious masses from normal tissue [7], See Figure 1. In recent years, Contrast Enhanced Spectral Mammography (CESM) has been gaining popularity in the screening of breast cancer due to its increased accuracy compared to DM, especially in people with dense breast tissue [8] [9]. CESM combines an injected iodinated contrast agent with standard mammography techniques to retrieve a low and high energy scan [9]. The low energy scan (ranging from 26-32 kVp [10]) is comparable to a regular digital DM scan, while the high energy scan (ranging from 45-49 kVp [10]) is combined with the low energy scan using a dual-energy subtraction technique to retrieve the contrast-enhanced version [11], see Figure 2 for the differences of the scans.

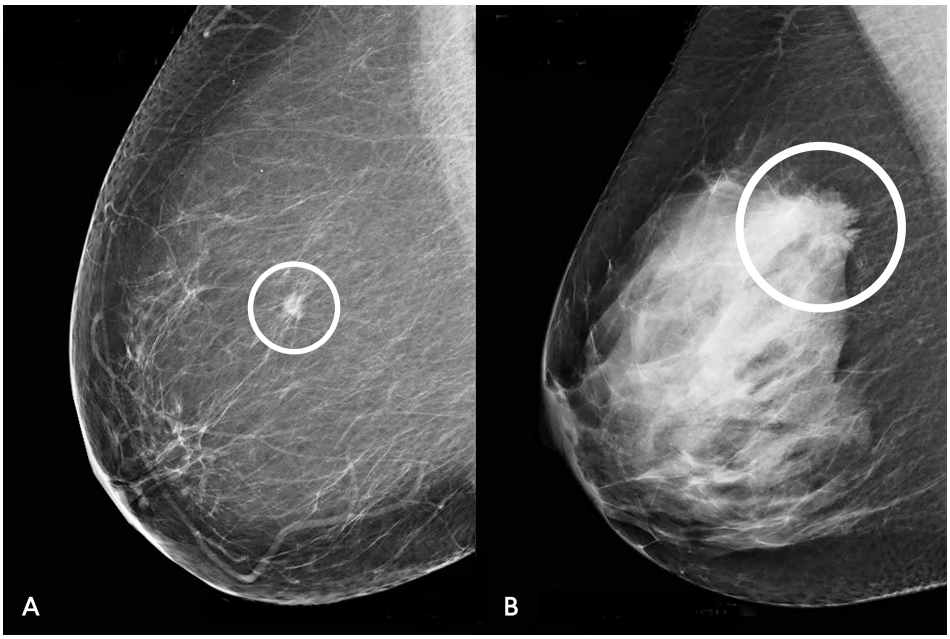


Figure 1: Comparison between identifying a tumour in a dense breast and non-dense breast, image taken from [12]

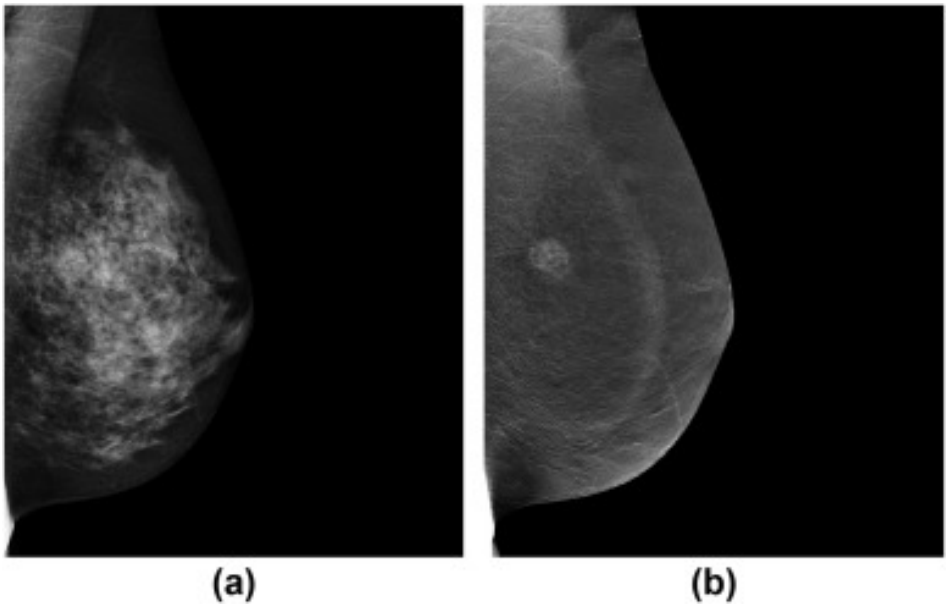


Figure 2: a) shows the output of the low energy mammogram image, the suspicious mass is not visible in this scan. b) Show the recombined image, where the suspicious mass is clearly present. Taken from [13]

Computer-aided diagnosis (CAD) refers to the use of computers and subsequent software to assist healthcare professionals in the analysis of medical images, such as MRIs

and mammograms, to diagnose diseases. The use of computers in the analysis of medical images is not a new concept, in fact, research on the topic exists as far back as the early 1960s [14] [15]. However, large-scale and systematic research and the development of different CAD schemes began 20 years later in the early 1980s [16]. This research ultimately led to the first FDA-approved CAD system to help diagnose breast cancer in screening mammography [17].

In recent decades, CAD systems have emerged as the primary way to help lessen the load on radiologists by serving as double-reading systems [18]. This is a system where two medical healthcare professionals individually analyse medical images before they come together to reach a consensus. While this is a valuable system that can drastically improve diagnosis accuracy, it is a system that can be incredibly time-consuming and resource-draining [19]. The use of CAD to act as the second opinion is therefore incredibly important, especially in organizations with a lack of radiologists such as the NHS, where a 2018 survey revealed that nearly 2000 more radiologists are needed to meet the level of demand for scans [20]. This understaffing of radiologists can result in longer wait times for patients and delays in diagnosis and treatment, highlighting the need for research to improve the performance of these CAD systems.

1.1 Objectives

The current objectives differ from the initial objectives set out in the DOER due to the project changing focus. The project aims are listed below:

1.1.1 Primary

1. To perform a literature review on the literature surrounding the use of deep learning in the segmentation of breast cancer.
2. To implement a basic U-NET model and a pipeline to segment breast cancer from CESM images, from the CDD-CESM dataset, which has not been used in context of a segmentation task in the literature.
3. To thoroughly investigate the effects of different individual affine data augmentation techniques on the performance of the segmentation model.

1.1.2 Secondary

1. To Perform a grid-search, with each of the best performing individual data augmentation technique, to find the best combination of data augmentation techniques for the CESM dataset.
2. To fine-tune the U-NET model, implemented using the grid search hyperparameter-tuning technique.

2 Literature Review

This literature review will explore the background and existing literature surrounding different techniques used in the CAD of mammograms. We first cover the use of different deep learning and non-deep learning techniques and architectures, following into the various publicly available datasets that are available to train these models, and then finally the use of transfer learning and data augmentation in the training of deep learning models in CAD.

2.1 Non-Deep Learning Techniques in CAD

Before the increase in the use of deep learning for image segmentation tasks, more traditional methods were used. This includes classical segmentation methods and machine learning models.

2.1.1 Classical Segmentation Methods

Classical segmentation methods such as edge and region-based segmentation rely on the fact that the pixel characteristics such as gray-level intensity, and texture in cancerous breast masses differ from non-cancerous pixel characteristics [21]. These differences are then used to distinguish between tumours and normal breast tissue.

Region-growing segmentation is a method where the pixels of an image are segmented into regions [22]. This is done by choosing a seed pixel, which is then grown into a region by grouping together neighbouring pixels that have similar pixel characteristics, which is determined by a threshold. The region stops growing when none of the neighbouring pixels are similar. This process is then repeated such that all the pixels in the image belong to some region [22]. In 2018, Punitha et al. improved upon the standard region growing technique by utilising the Dragon Fly algorithm, initially implemented by Mirjalili in 2015 [23], to generate optimal seed pixel placement and thresholds to better segment mammogram images from the DDSM dataset [24]. This algorithm emulates the static and dynamic behaviour of dragonflies for hunting and migration. The hunting behaviour is categorised by abrupt movements across a small area, while the migration behaviour is categorised by unidirectional movement along one direction [23]. This method outperforms standard region growing techniques because mammograms suffer from intensity variations between the different images, meaning that choosing a static threshold typically leads to poor segmentation [24]. They were able to achieve a Jaccard index of 0.90 using this novel method.

Watershed segmentation is a different region-based segmentation method that partitions the image into distinct regions with different elevations. Elevated regions correspond to higher pixel intensities, while depressions represent lower intensities [25]. Watershed segmentation simulates water filling from designated points, referred to as seeds, within the image. As the water flows, regions naturally form where water from

the different seeds meet [25], see figure 3. Ciecholewski explored the use of watershed segmentation on the segmentation of micro-calcifications from 220 ROIs taken from the DDSM dataset, 110 of which were benign and the other 110 were malignant [26]. The study used 4 stages of morphological transformations to detect the micro-calcifications in the mammograms and to locate their approximate location in the mammograms. After this, a watershed algorithm was used in order to extract the shapes of the micro-calcifications. With this method, Ciecholewski was able to achieve an IOU score of 0.708 which is impressive considering how small the micro-calcifications can be in mammograms. However, the study utilized ROIs instead of full mammogram images, meaning that an additional step (the extraction of ROIs) is required before new mammograms can be segmented.

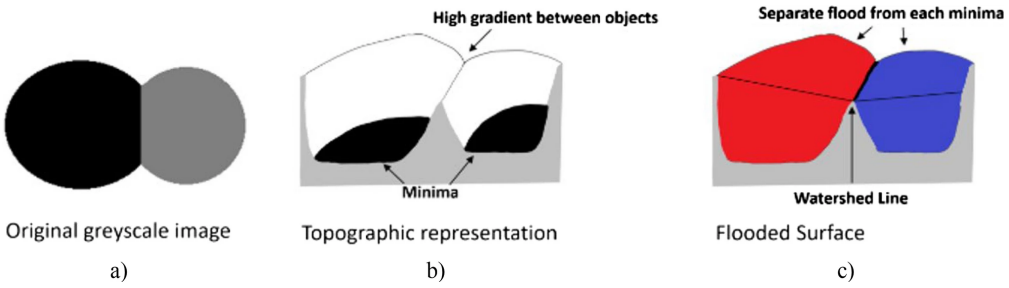


Figure 3: A diagram of the process of watershed segmentation from [27]. Image b) represents how darker pixel values have a deeper depressions.

While image segmentation in CAD is primarily used to segment cancerous masses, it can also be used in the pre-processing step to filter out structures that can act as noise. For example, the pectoral muscle is present in the Medio-Lateral Oblique (MLO) view in mammograms and can act as potential noise, as the density of the muscle is similar to the density of cancerous masses [28], see Figure 4. In 2012 Liu et al. utilized Otsu thresholding for the segmentation of the pectoral muscle from MLO view mammograms [29]. The Otsu threshold technique is an automatic segmentation technique which is used to segment greyscale images into regions using their pixel intensities [30]. The technique iteratively goes through all image intensities (0-255) and calculates the inter-class variances if it were to segment at that particular pixel value. The inter-class variance is the variance between the pixels in the two classes. Therefore, the segmentation threshold is chosen to be the one that maximises this inter-class variance [31]. In the study, the breast region is extracted from the background, after which the Otsu threshold method segments the pectoral muscle from the breast tissue. This segmentation was further refined using multiple regression analysis. This is a method where a polynomial of degree k is found which acts as a predicting function and has a minimum value for the sum of square errors. Unfortunately, the study does not compare the subsequent effect that the removal of the pectoral muscle has on the segmentation performance of the tumours in the dataset.

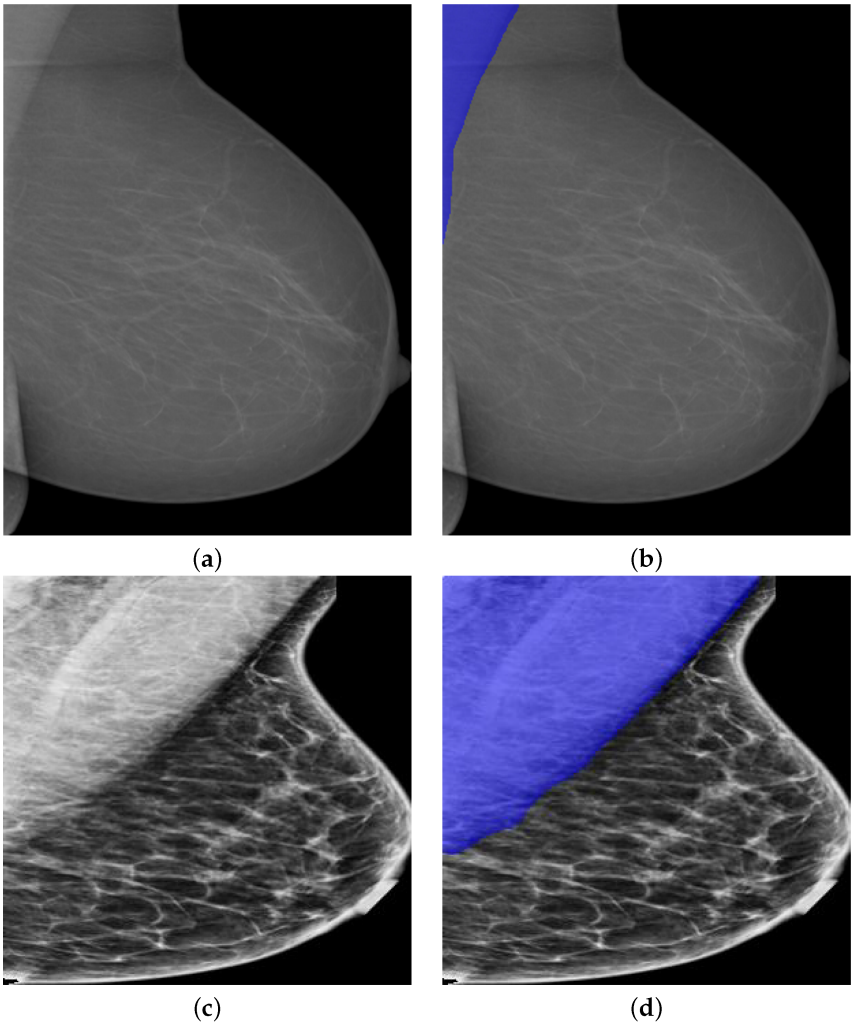


Figure 4: Segmentation of the pectoral muscles from 2 MLO view mammograms, taken from [32]

2.1.2 Machine Learning Methods

Prior to deep learning, other machine learning techniques were used in the CAD of breast cancer. The most popular machine-learning algorithms used were supervised learning where, the model is trained using annotated labels [33], which in the case of mammograms would be the classification of the breast cancer present for a classification task, or an annotated region of the cancerous mass for a segmentation task. An example of a supervised learning method is the k-nearest neighbour (kNN) algorithm, in which the model finds the k closest data points in the training set when seeing a new data point. The algorithm uses their labels to make predictions on unseen data. In 2018 Sharma et al. compared the accuracy of 3 supervised machine learning techniques in the classification of breast cancer from mammograms [34]. The three algorithms

compared were kNN, Naive Bayes, and random forest, and they were trained on the Wisconsin Diagnosis Breast Cancer dataset with 569 mammograms each labelled with one of two classes: benign or malignant. The study showed that while all 3 algorithms achieved fairly similar scores in different metrics, overall kNN performed the best, with a classification accuracy of 95.90%, and had the added benefit of having the shortest training time. Unfortunately, the paper is unclear exactly what features were extracted from the dataset and which was used for the 3 different models in the paper, thereby making it more difficult to replicate their results.

Unsupervised training occurs when the model learns from unlabelled data, this is done by finding patterns, or relationships within the data on its own. An example of this is Fuzzy C-Means (FCM) clustering, where a dataset is partitioned into C clusters in the context of image segmentation, and each pixel is assigned to one or more clusters based on its intensity value. However, unlike other clustering algorithms, each pixel is given a probability of belonging to multiple classes, this is referred to as membership [35]. This process is repeated with different thresholds until the sum of the squared distances between each data point and the centre of the cluster they belong to is minimized. Saleck et al. explored the use of FCM clustering for the segmentation of tumours in mammograms on images taken from the MIAS dataset [36], see figure 5. Saleck et al. combined the use of FCM clustering, pre-processing techniques such as Contrast Limited Adaptive Histogram Equalisation (CLAHE), and a feature extraction technique Gray-Level Co-Occurrence Matrix (GLCM) to extract 5 different features (contrast, energy, homogeneity, entropy and correlation) in the mass area to obtain an optimal threshold. Through this method, they were able to achieve an overall segmentation accuracy of 94.6%, a specificity of 96.4%, and a sensitivity of 86.2%. Although these metrics are quite impressive, they only tested on an 18-image sample of the dataset, which raises concerns about the generalisability of the results of the study. In addition to that, the study did not include metrics that are typically associated with image segmentation tasks such as Dice score or Jaccard index, making it harder to compare its results with other studies.

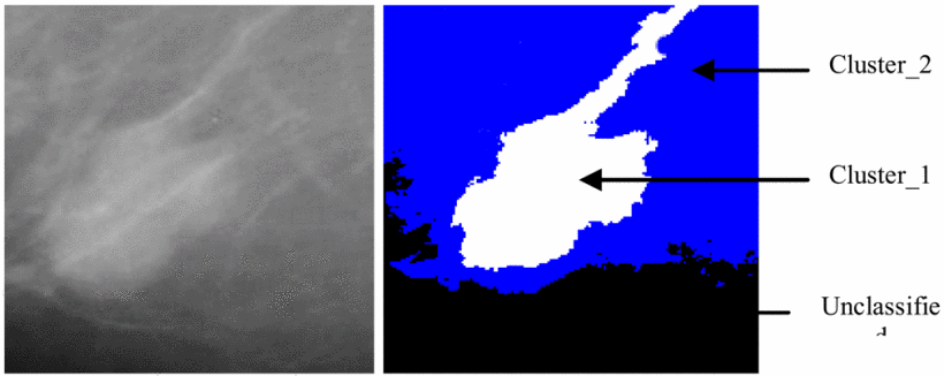


Figure 5: Segmentation performed by Fuzzy C-means clustering, taken from [36]

A downside to many of these machine learning algorithms is that they are unable to operate effectively on raw data, due to requiring features from the images such as Bare Nuclei, Clump thickness, and cell size to be extracted which is then passed into the algorithm [37]. This means that the performance and accuracy of these machine learning models are also dependent on the quality of the features extracted and whether these features can determine the regions of interest.

2.2 The use of Deep Learning CAD

In recent years, the research around CAD systems has been performed through the use of deep learning techniques. Deep learning is a subset of machine learning that uses deep neural networks, which are inspired by the structure of the brain, to process and learn from a large dataset. The rise of the use of deep learning can be attributed to two main factors: an increase in overall computing power, and an increase in the availability of public medical datasets [38], like the ones present in the Cancer Imaging Archive [39]. The main type of deep neural networks implemented in medical imaging are deep convolutional neural networks (CNN) because feature engineering is not required [40]. This allows the raw and unedited mammogram images to be passed directly into the models, making it significantly easier for radiologists to use. This simplicity could increase the uptake of CAD systems for use in hospitals.

2.2.1 Convolutional Neural Networks

CNNs are deep learning algorithms which are especially suited for image and pattern recognition. They are comprised of 3 main layers: Convolutional Layers, Pooling Layers, and fully connected layers.

The convolutional layers of a CNN focus on the extraction of features in an image, using filters called kernels [41]. These kernels are small matrices with a depth equal to the depth of the image, which is typically 3 for RGB. The kernel starts at the top left of the image and performs a dot product between the kernel and the patch of the image the kernel is highlighted over, see figure 6. After this dot product is calculated, the kernel moves a certain number of pixels to the right, which is defined by the stride that we have chosen [41]. Once the kernel has traversed the entire image, it will have a matrix called a feature map, which is then either passed through another convolutional layer or a pooling layer. As the model learns, the weights within the kernel change iteratively through the training process to extract features from the image that minimize the chosen loss function. Through this process, the kernels at the different layers of the model can extract different features from the image, which it can then use to perform different tasks, such as classification or segmentation.

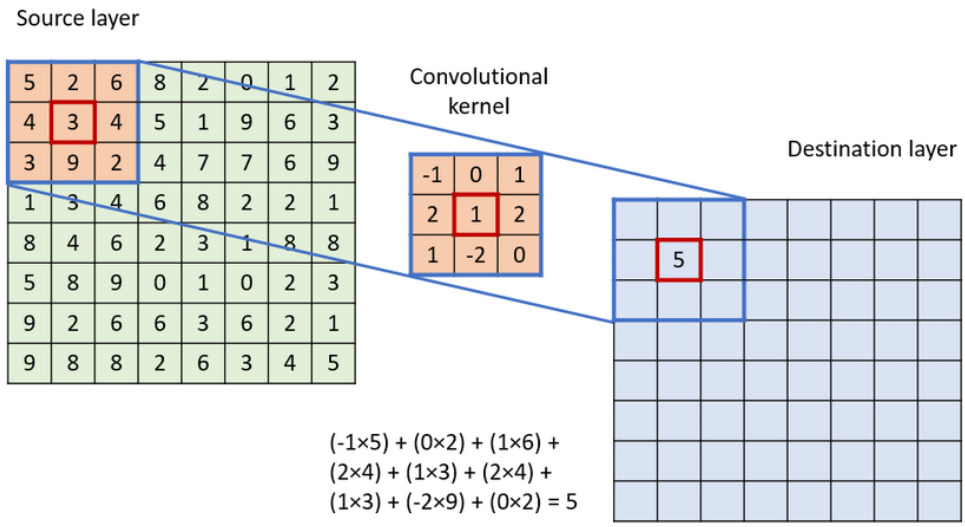


Figure 6: A graphic of the convolutional process on a 8x8 image with a 3x3 kernel, taken from [42].

Pooling layers reduce the spatial size of the feature map, decreasing the number of parameters in the model [41]. This reduction leads to a decrease in the computational power required to process the data and a reduction in the training time of the model. There are two commonly used types of pooling: max pooling, and average pooling. Max pooling is performed by looking at a small patch of the feature map, for example, a 2x2 section, and choosing the highest value, see figure 7. Conversely, Average pooling calculates the average of a patch of the feature map. In the case of detecting breast cancer from a mammogram, max pooling is typically more desirable, because it can extract dominant features that are present in the feature map. This is incredibly important as the tumours present in mammograms may only take a very small portion of the mammogram compared to the background and normal breast tissue.

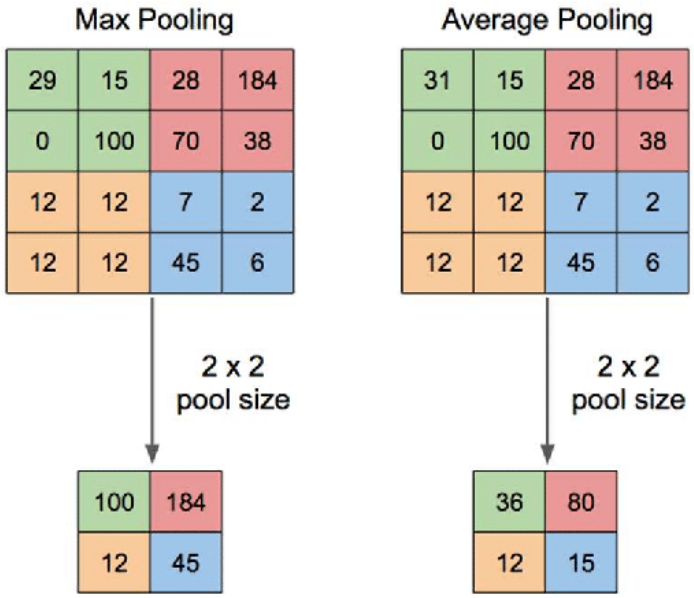


Figure 7: Effect of Max pooling and Average Pooling on a feature map. This is a 2x2 matrix with a stride of 2. Taken from [43]

The final layer in a typical CNN is the fully connected layer (FCL), a dense neural network (NN) where each neuron is connected to every neuron in the neighbouring layers. This layer attempts to analyse the features extracted by the CNN to perform a classification task by assigning a probability for each label. The backpropagation step of a CNN uses the errors calculated in the FCL to update the weights of all layers, including the weights of the kernels in the convolutional layers.

2.2.2 U-NET Architecture

A very popular CNN that is used for the segmentation of images is the U-NET model created by Ronneberger et al. in 2015 to win the ISBI cell tracking challenge [44]. This is a challenge where a segmentation model must be made to outline cells captured through a light microscope. The model achieved the best performance in the competition on two datasets, outperforming the second-best model by 9% in one dataset and then 31% in another dataset, despite only working on a dataset with 35 annotated images [44]. The model is an encoder-decoder network characterized by its U-shaped design, see figure 8. The encoder path is able to extract features about an image through the repeated use of two 3x3 convolutional layers with stride 1, followed by a 2x2 max pooling layer with stride 2. The decoder path up-samples the image through the use of a 2x2 up-convolution layer followed by a concatenation with the feature map on the same level on the encoding path. Unlike typical CNNs, the network does not contain an FCL, as it is primarily used for segmentation tasks. Instead, the final output layer of a U-NET consists of a single convolutional layer with a kernel of size 1x1. This layer performs

the final convolutional step to reduce the number of channels from 64 to the desired number of output classes.

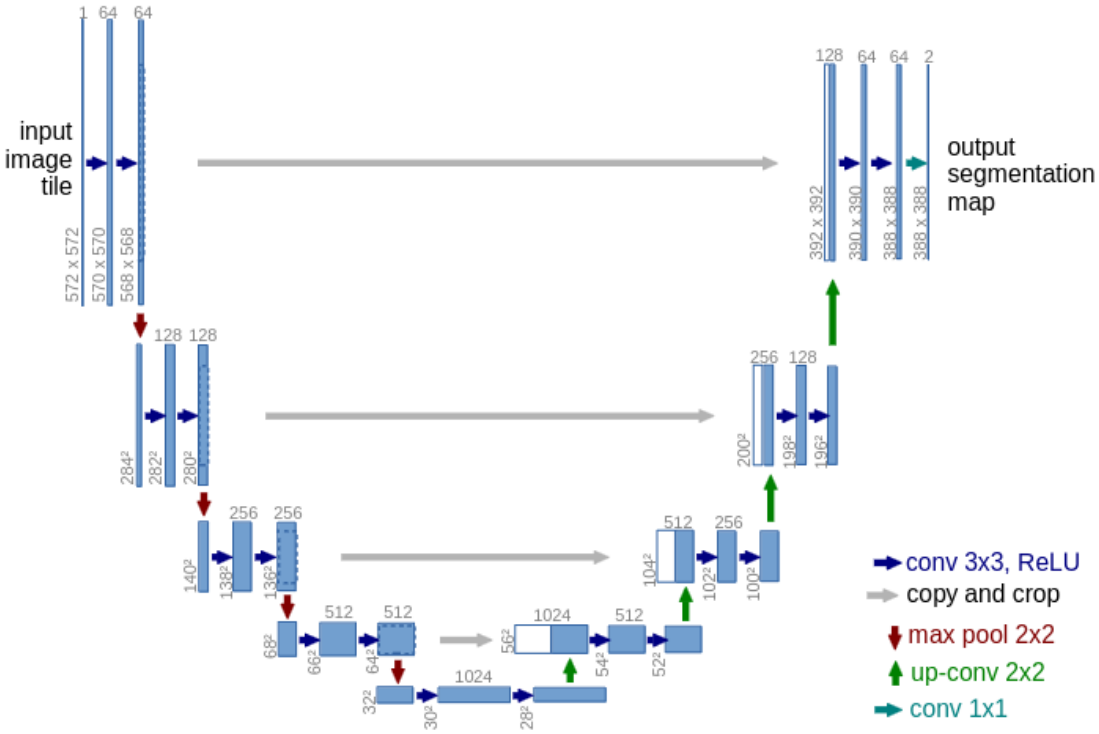


Figure 8: A diagram of the U-NET architecture from Ronneberger et al.'s paper. [44]

2.2.3 Deep learning in the Segmentation of breast cancer

After Ronneberger et al. published their paper on U-NET, its use in the segmentation of breast cancer in mammograms was widespread, due to its simplistic architecture and high performance. An example of this would be the 2020 study by Zeiser et al. which utilized the original U-NET architecture to segment digital mammograms from the popular DDSM dataset [45]. The study evaluated the effects of a different number of layers utilized in U-NET on the accuracy of a model, as well as the effects of data augmentation on ROI images. The study's best model had a specificity of 80.47% and a dice index of 79.39%. While these values seem inadequate to be used in practice as 1 in 5 people will be wrongfully called back for further testing, it is still impressive considering that there were an equal number of normal images and cancerous images used in the training, validation, and testing sets. This is uncommon as many studies in the literature specifically choose to only use images that contain masses, thereby making them inappropriate for use in a screening context, where a majority of the mammograms will not contain any masses. That being said, Zeiser et al. fail to justify the reason for their choice of hyperparameters for the model. For example, investigations into different settings for each of the hyperparameters could have been performed, or simply

citing papers in the literature that share the same settings would have been a simple addition to prove that the choices of hyperparameters were suitable.

Abdelhafiz et al. looked to improve upon the U-NET architecture for the segmentation of breast cancer through altering the operations performed in each layer of the encoding path [46]. Two main additions were added to each convolutional layer: a Batch Normalization (BN) layer and a drop-out layer. BN layers allow for faster training and increased generalisation accuracy of deep learning models by normalizing the input to a zero-mean and a constant standard deviation, allowing a higher learning rate to be implemented without the problem of potentially missing sharp local minima [47]. Dropout layers are a simple way to prevent overfitting in neural networks, a phenomenon where the model learns from the training data too well, and is unable to generalise its predictions to unseen data. Dropout layers have a probability of temporarily removing a *unit* from the network, which includes all incoming and outgoing connections, this prevents units from *co-adapting* decreasing overfitting [48]. With these changes, the study was able to achieve an impressive dice index of 0.951 and a mean IOU of 0.909 on a dataset of images collected from 3 public datasets CBIS-DDSM, INbreast, BCDR-01, and one private dataset UCHCDM. Unfortunately, like many other papers in the literature, the study only considers mammograms that contain masses, making it challenging to generalise the results to real-life clinical settings, thus limiting the findings of the paper.

Another issue with the U-NET model and other CNN models is how large the models are, for example, the U-NET consists of 31 million parameters [49]. This size can result in several drawbacks, including an increased demand for computational resources, leading to longer training times and higher hardware requirements. Moreover, larger models have a higher risk of overfitting, particularly when used with smaller datasets. Lu et al. investigated methods that can be used to drastically reduce the number of parameters of the standard of the U-NET model. Through experimentation, they verified that the benefits of feature fusion that are performed in the skip connections are minimal compared to the benefits of the "divide and conquer" strategy in U-NET's encoder. In the proposed Half-UNet, shown in figure 9, the channel numbers are unified such that it does not double after each downsampling step, this removes the redundant convolutions in place in the decoder path of the original U-NET - drastically reducing the model's parameters. Additionally, the standard double convolution is replaced with the ghost module, as introduced by Han et al. [50]. The ghost module aims to address the redundancy of intermediate feature maps calculated in CNNs through 2 separate stages. The first stage involves performing a controlled number of ordinary convolutions [50]. The second stage involves the creation of s ghost feature maps through a series of simple linear operations [50]. The model's performance was evaluated on the DDSM dataset, where it achieved a dice score of 0.8892, compared to the original U-NET model's 0.8939. The final Half-UNet model saw a negligible performance decrease while reducing the number of parameters from 31 million to 0.21 million.

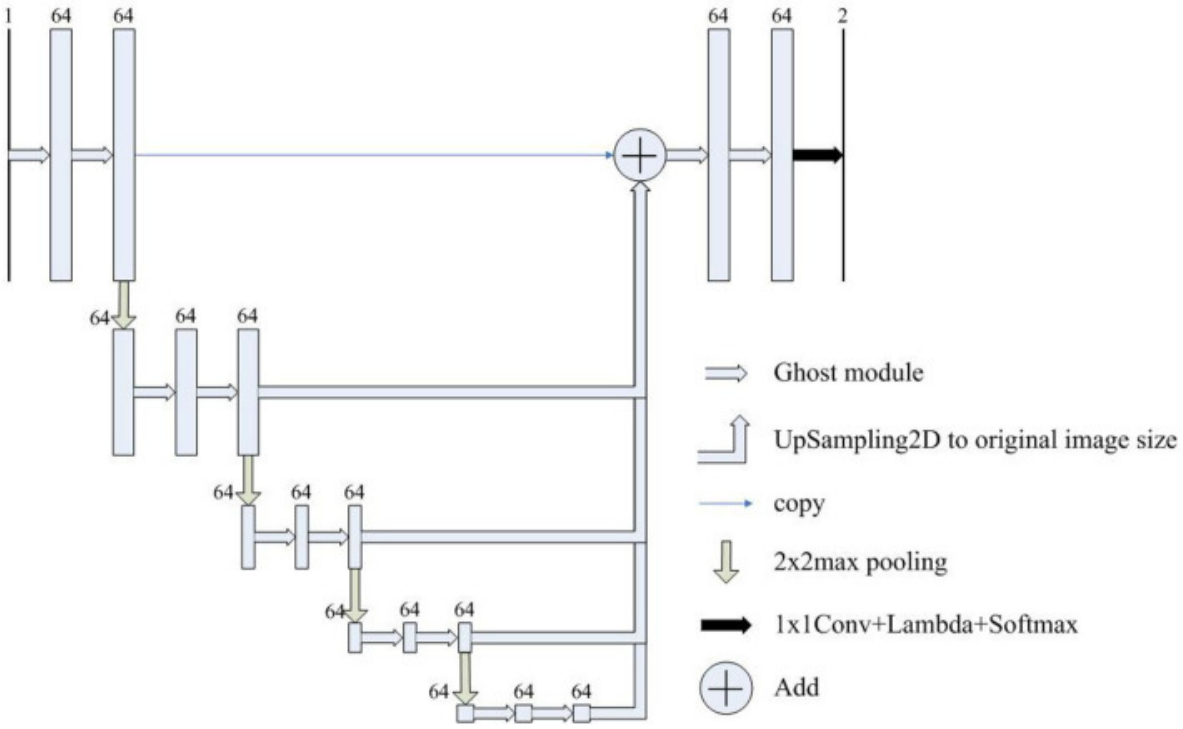


Figure 9: A diagram of the proposed Half-UNet from paper [49]. We see that the channels do not double at every downsampling step. Moreover, we see that the feature fusion step is performed by element-wise addition of the feature maps.

Another extension performed on the original U-NET architecture is its extension to be compatible with 3D image modalities such as MRI scans, as introduced by Cicek et al. in 2016 [51]. The overall architecture of the model is the same, except they have been adapted for 3D use by extending all operations to three dimensions, see 10. With this method they achieved a mean IOU of 86.3% in the segmentation of Xenopus kidney embryos, very similar to the performance that the traditional U-NET has [51]. In 2021, Hirsch et al. implemented 3D U-NET in the segmentation of breast cancer from breast MRIs and achieved an impressive median dice score of 0.77 which they found was, on average, higher than that achieved by radiologists which ranged from 0.69-0.84 [52]. They used a dataset of 64063 breast MRIs 60108 were benign and 3955 were malignant. They split the data into 3 separate sets: all the benign MRIs and 2455 malignant MRIs were used for training, 100 malignant MRIs were used for validation, and 250 malignant MRIs were used for testing. This split means that while their model was being trained on mostly benign tumours, the testing and validation set was solely consisting of malignant tumours, this class imbalance can be remedied by using data augmentation on the malignant cases to increase their number to be equal to that of the benign cases. After which, an equal number of benign and malignant cases could have been split into the testing and validation sets.

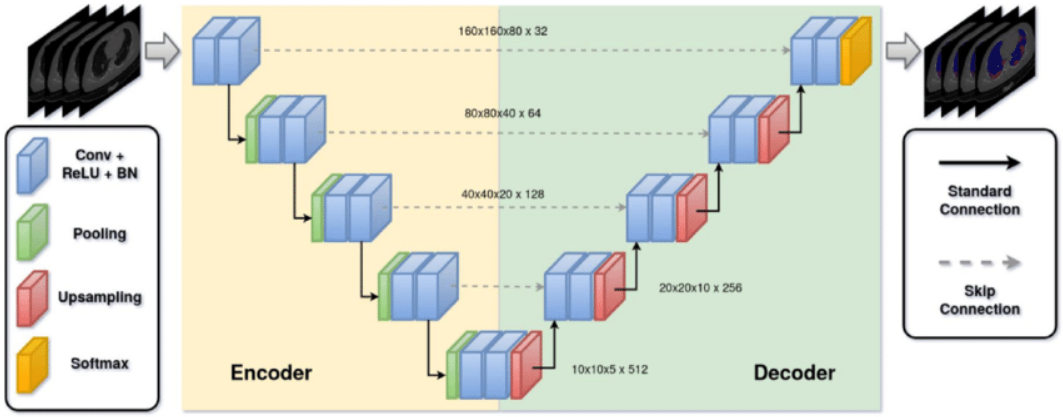


Figure 10: The 3D U-NET diagram used for the segmentation of Covid-19 infected lungs from CT images. Taken from [53]

While the segmentation of breast cancer is useful in the context of double reading systems, the automated segmentation of breast cancer can be used in pipelines for the detection, classification, and segmentation of breast cancer. For example, a 2018 paper by Al-Antari et al. utilised a full-resolution convolutional network (FrCN) to segment mammograms in the INbreast dataset [54]. The full pipeline consisted of three separate stages (see 11). A You-Only-Look-Once (YOLO) architecture was used to detect masses from the breast tissue, effectively retrieving mass ROIs. A newly proposed convolutional network FrCN was used, which allows the removal of the max-pooling and subsampling layers from the encoder network to preserve the spatial resolution of the original image, preventing any information loss that may have occurred during these layers [54]. Finally, a simplified version of AlexNet classified the segmented masses as either benign or malignant. With this pipeline, they were able to achieve an accuracy of 0.95, a segmentation Dice index of 0.92, and an IOU of 0.86 on the INbreast dataset. Unfortunately, similar to [46], the mammograms only contained instances of benign and malignant cases.

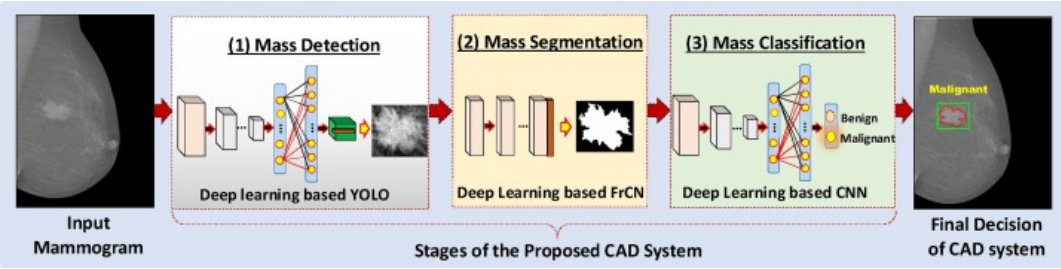


Figure 11: The pipeline used by Al-Antari for the detection, classification, and segmentation of breast cancer. Taken from [54]

Alternatively, instead of using different models to perform the tasks of detection, seg-

mentation, and classification, Soulami et al. utilised the original U-NET model for all 3 stages of the pipeline [55]. In this study, the greyscale digital mammograms are converted to RGB as this can enhance the dynamic range of the image, making it easier to spot the difference between normal breast tissue and a cancerous mass [56]. In addition to this, the segmentation masks are coloured to indicate which of the 3 classes a pixel belongs to red being normal breast tissue or the background, green representing a benign mass, and blue representing a malignant mass. The classification of a mass is performed by observing which colour is present in the predicted mask. The benefit of using solely the U-NET model to implement a detection, classification, and segmentation pipeline of breast cancer is that the performance of all 3 stages is solely reliant on the performance of the U-NET model. This is in contrast with [54], where the performance of the segmentation is reliant on the performance of the detection, and the performance of the classification is reliant on the performance of the segmentation. With this, they achieved a segmentation IOU of 90.50% and an impressive 99.19% dice index, on 1079 images collected from the DDSM, CBIS-DDSM, and Inbreast datasets. These results are impressive considering that the combined dataset used contained an equal spread of the 3 classes, unlike [46] and [54]. However, it is essential to acknowledge potential drawbacks in the study. One is the lack of justification on which images were chosen from each of the datasets. Without clear criteria for the image selection process, the study's sample may not accurately represent the diversity of cases within each dataset, potentially impacting the validity of the results.

2.3 Publicly Available Datasets

One limitation associated with deep learning is its reliance on a substantial amount of data to achieve optimal performance. This is especially pertinent in the context of medical imaging, as possible inaccuracies could mean drastic consequences for patients. As it may lead to potential cancers being missed or unnecessary biopsies being performed.

The most common dataset used in studies is the Digital Database for Screening Mammography (DDSM) or a subset of it, the Curated Breast Imaging Subset of DDSM (CBIS-DDSM). DDSM is a database of mammograms that was created in 1997 and contains 2620 scanned film mammography studies, each consisting of between 6 and 10 files: a file providing information about the case as a whole, four image files, an overview file containing all four images, and between 0-4 overlay files. There are plenty of issues with DDSM such as the fact that images are saved in non-standard compression files which necessitates the use of out-dated decompression code, and that the ROI annotations indicate the general position of suspicious masses rather than precise outlines [57]. The CBIS-DDSM dataset is a subset of DDSM containing 1556 patient scans that aims to rectify these issues by converting the images to the DICOM format, which is widely used in medical imaging [58]. As well as that, the ROI segmentation has been made more precise by a trained radiologist [57].

Two more very common datasets that are used are the MIAS (Mammographic Image

Analysis Society) and the INbreast dataset [59] [60]. These datasets are smaller than DDSM, only containing 322 and 410 images, respectively. While these datasets were widely used a number of years ago, their small sizes and the advent of newer and larger datasets mean that these datasets are less used now, or are used in conjunction with other datasets as seen in [55] and [46].

A new dataset that was published in 2021 is the Categorized Digital Database for Low energy and Subtracted Contrast Enhanced Spectral Mammography Images (CDD-CESM). This is the first public dataset that provides data for CESM mammography [1]. The dataset contains data on 326 patients, each consisting of the low energy and contrast enhanced versions of the mammograms, which includes the clinical data of the patient in the form of a BI-RADS classification of the breast and segmentations of ROIs. This is the dataset chosen for this study and is expanded upon in Section 4.2.

While there has certainly been a drastic increase in the number of mammogram datasets that are now present online, it is still not enough to get the most out of the deep learning models. Many studies perform techniques such as transfer learning which initializes the weights of the neurons after being trained on a large image dataset, or data augmentation which can artificially increase the size of datasets through image transformations.

Dataset	Number of Patients	Number of Images	Date of Release	Masks
DDSM	2620	11560	1997	Yes
CBIS-DDSM [57]	1566	6671	2017	Yes
CDD-CESM [1]	326	2006	2021	Yes
INBreast [60]	115	410	2012	Yes
MIAS [59]	161	322	1994	Yes, Elliptical

Table 1: Information on the public datasets identified in the literature review that were used in a segmentation context.

2.4 Transfer Learning and Image Size in Digital Mammograms

A popular technique that researchers have investigated is transfer learning, where the model is first pre-trained on a large image dataset. After which, the pre-trained weights are then fine-tuned by training the model on the desired dataset. Adam Jaamour et al. have shown that using any kind of transfer learning (on all layer weights/fully connected layer weights/ImageNet weights only) produces more accurate models in the classification of breast cancer as opposed to initializing the model with random weights [61]. The study found that the use of transfer learning improved the model's accuracy by 5.6%.

A popular image dataset used in many applications of transfer learning in deep learning models is ImageNet [62]. ImageNet is a large and varied image dataset with classifications of different objects see figure 12 for examples of images present in the dataset.

Pre-training models on these large datasets allow the models to initialize their weights, effectively transferring what they have learned onto a new dataset [63]. This bypasses the need for the model to learn lower-level abilities such as edge detection, thereby decreasing training time. A way to incorporate these pre-trained models into a U-NET architecture is to use some other model, such as VGG16, as the encoder for U-NET [64].

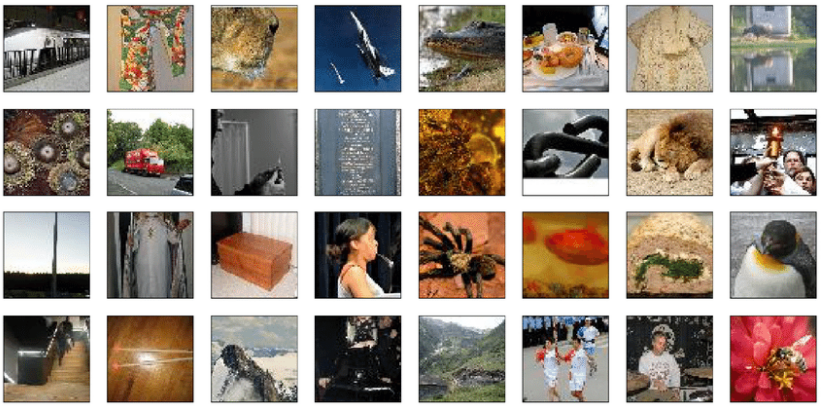


Figure 12: Examples of images present in the ImageNet database. Taken from [65]

A downside to using transfer learning with datasets such as ImageNet is that to maximise the effectiveness of the final weights assigned, images must be resized to be smaller. The typical image sizes used are 224x224, 256x256, and 512x512. These image sizes are much smaller than the images of mammograms you would find on datasets such as DDSM. This can be detrimental to the accuracy of a model, as breast cancer can be difficult to detect even in full mammograms, especially in women with particularly dense breast tissue [7], therefore reducing the size of the image may lead the model missing potential tumours. A 2022 study by Ranjbarzadeh et al. trained a multi-route feature extraction model to segment patches of mammograms [66]. This is a technique in which the CNN only processes one patch at a time, rather than the whole image, allowing us to circumvent the resizing of entire mammograms. Another example of avoiding the resizing of images is present in [45], where regions of interest (ROIs) were used as training data in conjunction with full mammograms.

2.5 Data Augmentation

Data augmentation is another technique used to improve the performance of a model. The first category of data augmentation techniques are called affine transformations. Affine transformations are ones which preserve parallel lines after they have been performed onto an image. Examples of affine transformations are: flipping, rotation, zooming, shearing, and translation. These are transformations that can easily be added to a dataset to artificially increase the number of images used to train a model. Studies have shown that the addition of these augmented images in a dataset can decrease the

likelihood of a deep learning model from overfitting [67]. Many studies that perform research on the CAD of breast cancer often use data augmentation to increase the number of images that used to train the model, see Table 2. However, many of these studies fail to evaluate the performance increase that data augmentation has had, and some even fail to mention the exact data augmentations used on their training data. None of the studies listed evaluate the effect that different data augmentations can have on their model, which may lead to potentially superfluous images being added to the training data which do not have any effect in the training model (thereby increasing training time for no performance increase) or may even be negatively effecting the model by learning behaviours that do not exist in an actual mammogram, for example a rotation of 90°.

Paper	Dataset	Rotation	Cropping	Zooming	Flipping	Translation	Improvement	Final Metric
[46]	INbreast, CBIS, BCDR-01	✓	✗	✓	✓	✓	Dice score from 0.922 to 0.951	Dice score of 0.951
[45]	DDSM	✗	✗	✓	✓	✗	Increases of 0.1 and 0.2 in validation dice score	Testing Dice Score of 0.79
[49]	DDSM	✓	✗	✗	✓	✗	N/A	Dice Score of 0.8944
[54]	INbreast	✓	✗	✗	✗	✗	N/A	Dice Score of 0.9269
[68]	DDSM	✓	✗	✗	✓	✗	N/A	Dice Score of 0.8224
[69]	CBIS-DDSM	✓	✗	✗	✓	✗	N/A	Dice score of 0.9475
[70]	MIAS, DDSM, CBIS-DDSM	✓	✗	✗	✗	✗	N/A	Dice Score of 0.9189
[71]	MIAS, CBIS-DDSM, IN-breast	✗	✓	✓	✓	✗	N/A	Dice Score of 0.9722
[72]	INbreast	✗	✗	✗	✓	✗	N/A	Dice Score of 0.88

Table 2: A table of various data augmentation techniques used in the literature. Many fail to state the improvement, or lack thereof, that the addition of data augmentation made on their model.

If affine transformations do not increase the performance of a CNN, elastic transformations can be used. These transformations apply slight deformations to images to generate similar images. These transformations have to be carefully parameterised, such that the augmented images retain the same characteristics and properties as the original [73]. A popular technique of elastic deformation is called randomised displacement field and was first utilised by Simard et al. in 2003 on the MNIST dataset [74]. This technique involves the creation of a randomised displacement field, which shifts each pixel by a random value between -1 and +1 in the x and y direction. The field is then convolved with a Gaussian with standard deviation σ. This can be considered an elasticity coefficient, as small values of σ will result in a random field, whereas large values of σ will result in small displacements [74]. In addition, the intensity of the displacement is altered by multiplying the displacement fields with a scaling factor α. Low values of α will result in random displacement, intermediate values of α will result in elastic transformations, whereas large values of α will result in transformations that are similar to translations [74]. Both σ and α can be considered hyperparameters, as they need to be finely tuned to the dataset at hand to ensure only images within the domain of the dataset are generated. Simard et al. Found that these elastic transformations achieved greater results on the MNIST dataset compared to using affine transformations. This could be because elastic transformations created images with greater variation in handwriting style, producing a more generalisable CNN.

Castro et al. applied this method to a collection of mammograms from 3 datasets: INbreast, CBIS-DDSM, and BCRP, to increase mass detection performance [75]. Each pixel was given a random value from the range -0.5 to +0.5, after which they were multiplied by the displacement scalar $\alpha = 300$ and convolved with Gaussian, with $\sigma = 20$. This field was then applied to the images, masks, and mass annotations. The result of this transformation is shown in figure 13. Through the addition of these elastic transformations, Castro et al. noticed a sharp decrease in the false positives per image for 2 out of the 3 datasets [75].

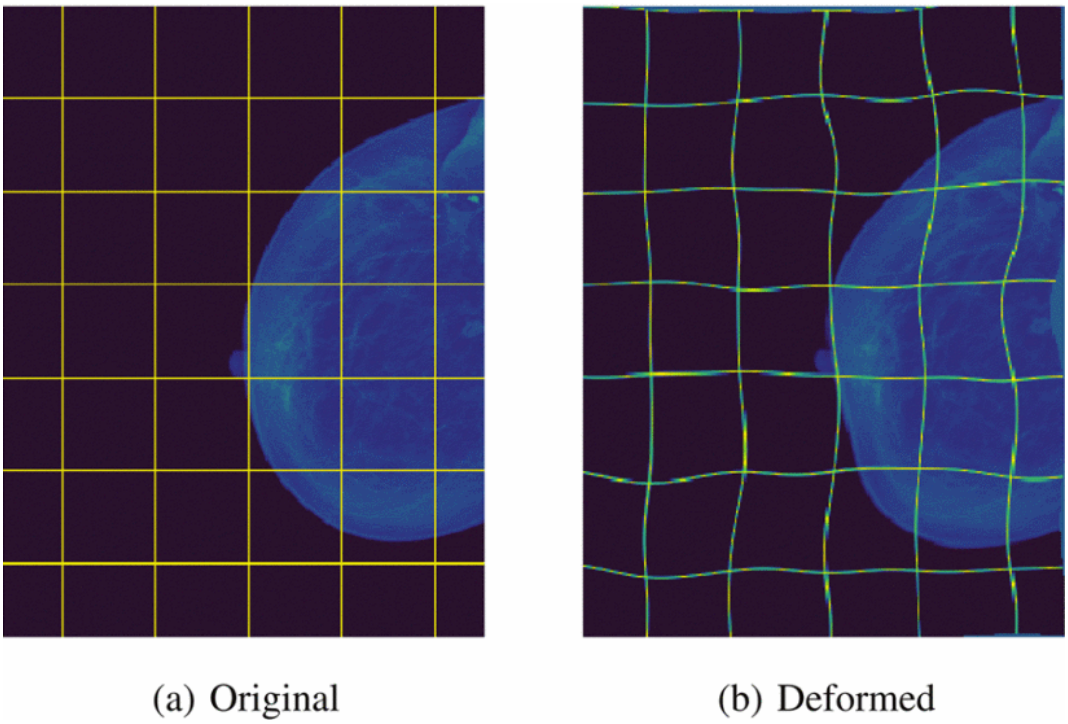


Figure 13: The result of random displacement field being applied to a mammogram. Taken from [75].

A more complicated form of data augmentation could involve the use of generative adversarial networks (GANs). GANs, a technique of generating synthetic data, were introduced by Goodfellow et al. in 2014 [76]. This method involves training two neural networks simultaneously, a generator and a discriminator. The generator is trained to generate data samples that resemble real data by taking in random noise which is used in the generation of data. The discriminator is trained to distinguish between real data and synthetically generated samples by the generator. This simultaneous training can be represented as a two-player minimax game, where G aims to minimise the value $log(1 - D(G(z)))$, which represents the likelihood of the discriminator guessing that a synthetic image is real, and D aims to maximise the value $log(D(x))$ which is the likelihood that discriminator guesses a real image is not synthetic [76]. Ever since this discovery, GANs has been a heavily researched topic in the field of data synthesis. This influx of research on the topic has led to the creation of a multitude of different GANs architectures, such as cGAN, DCGAN, and many more [77].

Haq et al. utilised a GANs architecture in the segmentation of breast cancer in MRI images from the RIDER dataset [78]. However, instead of being used for the generation of new data, it was used as a loss function to train the generator to segment the input images. The generator was created using an improved U-NET, which utilised a parallel dilated convolution (PDC) module instead of the regular bottleneck layer. This improves the ability for the model to recognise tumour related characteristics which may

be of different sizes without increasing the number of parameters of the model [78]. The output of the generator was a 256x256 prediction mask. The discriminator was a Patch-GAN deep CNN which classifies segments of the input image as being fake or real. The discriminator was convolutionally applied to the entire input, and then the responses were averaged and scaled to 0 (fake) to 1 (real). The input to the discriminator was a concatenation of the input/target pair in the dataset and the input/prediction pair generated by the generator. Through this method Haq et al. achieved a dice score of 0.85, much higher than other papers which had also used the same dataset.

2.6 Direction

Through the research conducted while writing the literature review for this dissertation, it became evident that the CDD-CESM dataset has not been utilised for a segmentation task in previous studies. Therefore, the primary direction of the dissertation is to provide a baseline assessment of the segmentation performance on the dataset. Additionally, while the advantages of using data augmentation techniques are acknowledged in the literature, a thorough investigation into the effect of different data augmentations on mammogram segmentation performance has not been performed. Therefore, this dissertation aims to thoroughly explore the effect of different data augmentation techniques (being flipping, rotation, shearing, and translation) have on the segmentation performance of a U-NET model.

3 Ethical Considerations

This project aims to investigate the effect of different data augmentation techniques in machine learning in the segmentation of breast cancer from mammogram images. The project will use the CDD-CESM dataset from The Cancer Imaging Archive. No data will be collected, and no surveys or interviews will be required.

The ethical issue raised through this dissertation is the use of a secondary dataset of patients' medical scans. This data has been anonymised and therefore contain no identifying information for the patients. Additionally, there will be no attempt to identify any of the individual patients to maintain confidentiality. The data will be stored on the encrypted Computer Science servers, and all the investigations will be run on a school GPU machine. The data is available for public access on The Cancer Imaging Archive [39]. This data will be used to train a Convolutional Neural Network in the segmentation of breast cancer from mammography images. To evaluate this model's performance, the dice score metric will be used. The methods used to work with the images will include data-cleaning, and the generation of new images using data augmentation techniques.

This work has received full ethical approval from the University of St Andrews, please see the written approval in Appendix 9.2.

4 Methodology

4.1 Approach

The first step of this dissertation was to implement a basic model (a basic U-NET with batch normalisation layers included) as well as a basic learning pipeline to gather baseline results on the segmentation performance on the dataset. At this point in time, the pipeline implemented did not include any data augmentation steps.

After this, a series of hyperparameter-tuning experiments, comprised of multiple grid searches of different parameters, were run on different configurations of the parameters. This step was performed to ensure that the U-NET implemented was being run with the optimal set of hyperparameters.

Furthermore, a comprehensive search is then performed across the different search spaces of the different individual affine transformations. This search allows us to assess the efficacy of the different data augmentation techniques on the performance of the model's ability in the segmentation of the breast cancer. Moreover, the optimal parameters for each of the augmentations can then be used in the next investigation, drastically reducing the search space of the following grid search.

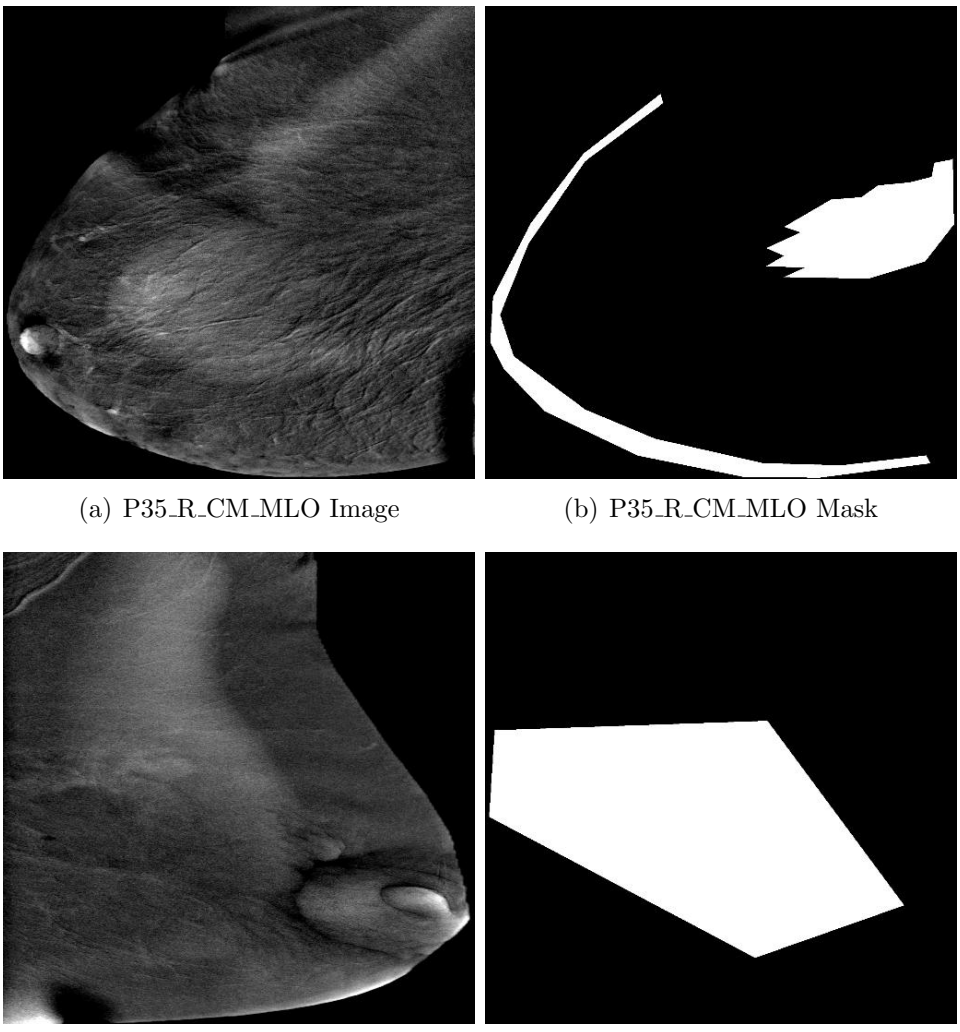
Finally, we explore the synergistic effect of combining the images produced from each of the individual transformations together in a grid search. In this grid search, we aim to find the best combination of data augmentation techniques that allow the model to generalise the best.

4.2 Dataset

The dataset used in this dissertation is the 2021 CDD-CESM, which contains contrast-enhanced and normal digital mammograms of 326 patients [1]. This dataset was downloaded from The Cancer Imaging Archive [39]. The dataset comes with the following resources:

- The CESM and DM images of the breast in both the MLO and CC views. This has already been converted from the DICOM format to JPEG using a lossless compression scheme. In addition to this, the images are cropped and zoomed to the breast region.
- A zip file containing the clinical data reports in the DOCX file format.
- A CSV containing the vertices of the hand drawn segmentations for each of the images.
- A CSV containing annotations for each of the images. For example, an image may have multiple annotations: suspicious, malignant, mass. These annotations allow us to quickly filter out images in a dataset if they do not contain annotations of interest.

The dataset contains a total of 2006 images, evenly divided between, 1003 contrast-enhanced images and an equivalent number of regular digital mammograms. Unfortunately, not all of these images could be used, this is because a substantial portion of the dataset contained images with annotations that were not associated with suspicious masses or calcifications. For example, over 100 images were labelled as being postoperative, meaning that the mass initially that was considered either benign or malignant may have been surgically removed. The addition of this data into our model may have been akin to adding more noise into our dataset. Moreover, a lot of the images that were not annotated with 'mass', 'suspicious', or 'calcification' had very strange masks as shown in figure 14.



(a) P35_R_CM_MLO Image (b) P35_R_CM_MLO Mask

(c) P44_L_CM_MLO Image (d) P44_L_CM_MLO Mask

Figure 14: Two different images with strange and large masks. Both were annotated with 'postoperative' in the annotation CSV.

Therefore, to avoid the issue of these images acting as noise in the dataset, images that did not contain any of the tags of interest were removed. This adjusted dataset had a total of 752 images, 414 images contained no masses, and 338 images were either benign or malignant.

This dataset was then split into a 70-15-15 training, validation, and testing split. This left us with 526 images in the training set, 112 images in the validation set and 114 images in the testing dataset. A distribution of the classes in each of the subsets are shown in table 3. The test set would remain unseen until all the investigations on the validation set were complete. This set was used to evaluate the true performance of the model on unseen data.

	Normal	Benign	Malignant
Train	289	44	193
Validation	60	10	42
Test	65	9	40

Table 3: The final distribution of the dataset in the three subsets: train, validation, test. Benign and Malignant classes were treated as being the same class.

4.3 Data Augmentations

The data augmentations investigated for this dissertation are affine, where parallel lines are preserved after the transformation. This was done as these are the easiest and quickest data augmentations that one can use to enlarge a dataset, without worrying about tinkering with hyperparameters. Some of the data augmentation techniques have the potential to produce images that were not initially present in the original dataset, such as a vertically flipped mammogram. However, this is still an important property to investigate, as these images may allow the model to generalise more, thereby increasing segmentation performance.

4.3.1 Flipping

Mirroring the image horizontally and vertically is perhaps the simplest data augmentation technique that can be used. This involves mirroring the image along either the x or y-axis. Flipping along the y-axis is very common as it simply flips the mammogram from being of the left breast to the right breast, or vice versa. Vertical flips on the other hand are more uncommon, due to the fact that the images they produce are not present in the original set of images.

4.3.2 Rotation

Rotation is another very common data augmentation technique to perform, as many mammograms exhibit slight orientation differences due to the specific machine used,

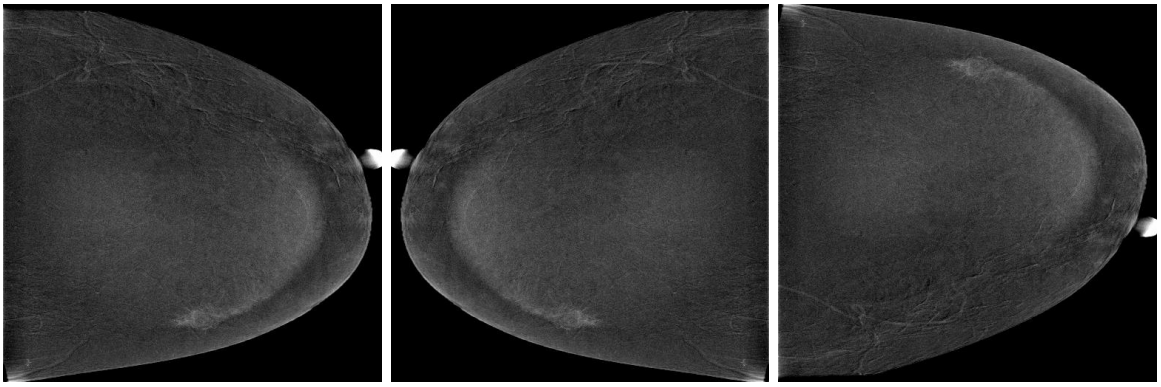
variations in breast characteristics, and positional nuances during the imaging process. Therefore, rotating images by a small amount will allow us to enlarge the dataset with similar images. Furthermore, rotations of larger angles, such as 45° and greater, are also investigated in this dissertation.

4.3.3 Shearing

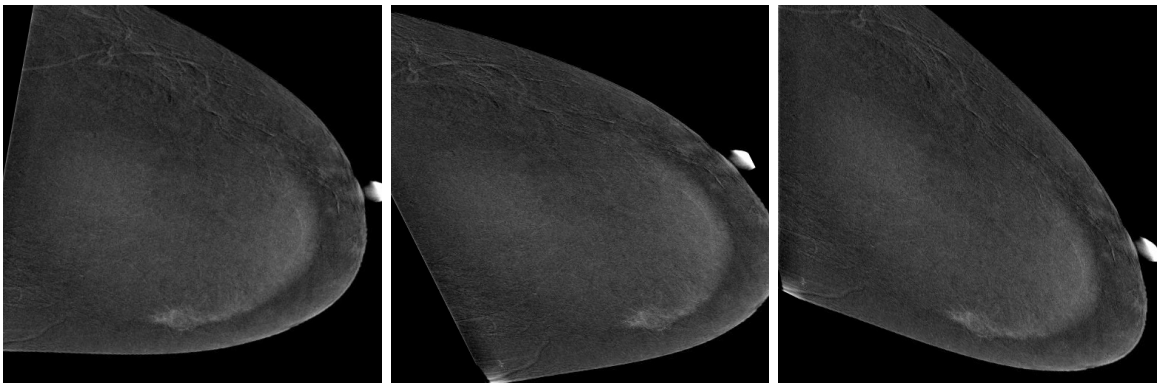
Shearing is a data augmentation technique that distorts the shape along one axis, while leaving the other axis unchanged. The result of a shearing operation is an image which has been distorted into a parallelogram-type shape. Much like rotation, shearing can be applied to address the slight orientation differences between the mammograms in the dataset.

4.3.4 Translation

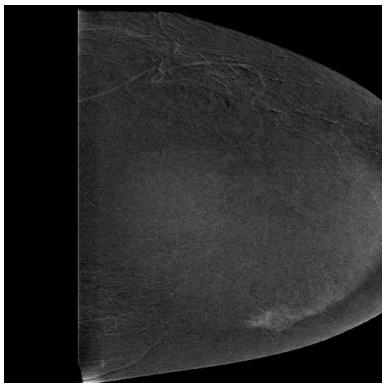
The translation of the mammogram along the x-axis was also investigated, even though it is a technique that is rarely used when working with mammograms. The translations performed were to move the breast region of interest closer to the centre of the image. This decreases the differences between the right and left breast mammograms and hopefully would lead to greater generalisability of the model.



(a) Original Image (b) Horizontally Flipped (c) Vertically Flipped



(d) Rotation of 10° to the left (e) X Shear of 0.5 (f) Y Shear of 0.5



(g) Translation of 100 pixels

Figure 15: The effects of the different data augmentation techniques on the file P_3_L_CM_CC.jpg image

4.4 Pipeline

Many studies that work on mammogram datasets typically have a very extensive pipeline due to the number of pre-processing steps that have to be performed on to

the digital mammograms. For example a majority of the studies that work with the DDSM dataset have to perform a variety of pre-processing steps:

- Decompression of the images using outdated decompression code.
- Contrast Limited Adaptive Histogram Equalisation (CLAHE): CLAHE is a technique to remove noise from photos and allows lesions to be more easily seen [79]. This is a technique that is prevalent in all research conducted on digital mammograms due to its effectiveness.
- Breast detection is usually performed to highlight the breast area present in the mammogram, this is done so that images can be cropped to remove a majority of the background from the image.

These steps did not have to be performed on the CESM dataset as the images have already been pre-processed. Therefore, the pipeline is very simple, with the only pre-processing step being the addition of the augmented images into the training set. The full pipeline is shown in figure 16.

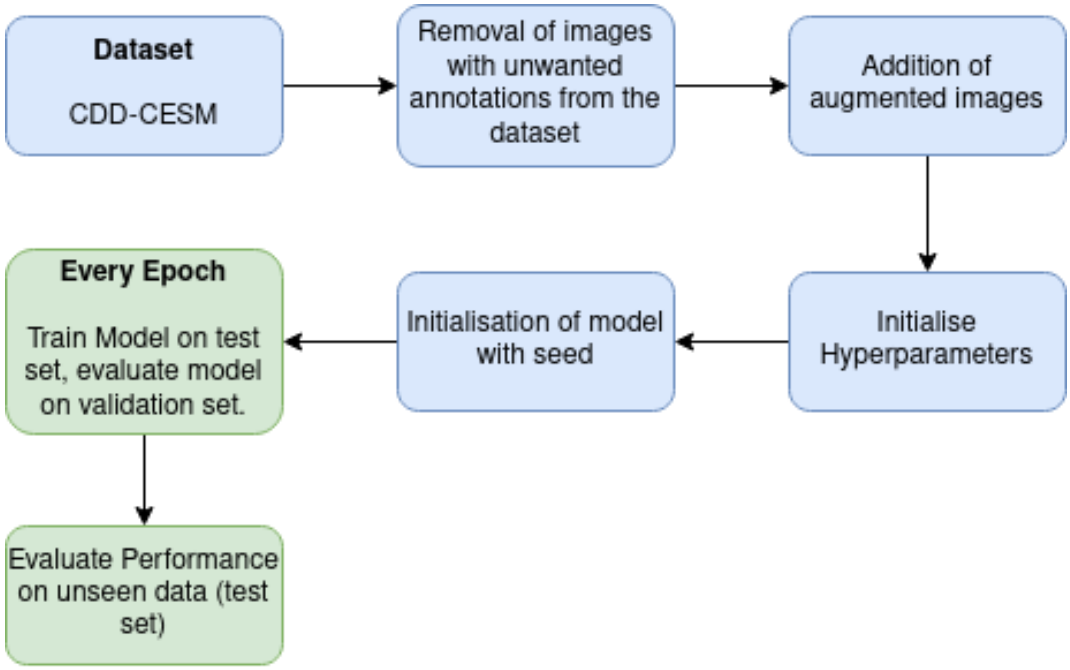


Figure 16: The pipeline used in all the investigations performed in this dissertation..

4.5 Technologies Used

The experiments were performed on a University GPU machine which contained an Nvidia RTX 3060 GPU with 12GB of VRAM, 32GB of RAM, and an Intel Core i5-6500 CPU. This machine was only accessible remotely via SSH, requiring key-based authentication. All the experimental results were stored on the GPU machine's 400GB

SSD. The model and pipeline implementation for this dissertation was implemented using the PyTorch library inside a python virtual environment. The data augmentations were implemented using the OpenCV, and Albumentation python libraries.

4.6 Performance Metrics

The main two performance metrics measured for the evaluation of the models on the validation dataset was the dice coefficient and accuracy. On top of these metrics the precision and recall were added to measure performance in the testing dataset. These equations will utilise the following abbreviations:

- TP (True Positive) - When a pixel in the prediction and the corresponding pixel in the ground truth both equal 1 (suspicious mass present).
- TN (True Negative) - When a pixel in the prediction and the corresponding pixel in the ground truth both equal 0 (normal tissue/background).
- FP (False Positive) - When a pixel in the prediction is 1, but the corresponding pixel in the ground truth equals 0.
- FN (False Negative) - When a pixel in the prediction is 0, but the corresponding pixel in the ground truth equals 1.

4.6.1 Accuracy

The accuracy metric measures the number of correct pixel predictions out of all the pixels in each of the prediction masks.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FN + FP)} \tag{1}$$

4.6.2 Dice Coefficient

The dice coefficient is a very popular metric when reporting on the performance of a segmentation model. It measures how much of the prediction mask overlaps with the ground truth mask. In order to avoid having a 0 denominator, a very small smoothing factor ϵ is added to the numerator and denominator. The dice coefficient of two identical binary images is equal to 1, and the dice coefficient of a binary image with no overlap is 0. The dice coefficient can still work well in datasets where the average sizes of the classes differ significantly, as it only measures the overlap between the two positive classes. This is therefore not only a useful metric in the context of breast cancer, as masses may only take a very small portion of the image, but also in medical imaging as a whole.

$$Dice = \frac{(2 \times TP + \epsilon)}{TP + FP + TP + FN + \epsilon} \tag{2}$$

4.6.3 Precision

Precision measures the proportion of how many of the positive pixel predictions were actually correct.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

4.6.4 Recall

Recall Measures what proportion of the pixels that actually had a value of 1 were predicted as having a 1 in the prediction mask. This is a useful metric to have as it can indicate whether the model was underpredicting, which can happen with datasets that have a higher proportion of 'normal' classes, like the filtered CDD-CESM dataset we are using.

$$Precision = \frac{TP}{TP + FN} \tag{4}$$

5 Validation Set Investigations

This section outlines the various findings of the investigations performed on the validation dataset results. The investigations covered are: implementation of a U-NET model in Section 5.1, a grid search to find the ideal hyperparameters in Section 5.2, a grid search to find the best lost function and optimiser pairing in Section 5.3. After these three investigations the model will be very finely tuned, a seed search will then be performed in Section 5.4 to ensure the model starts with favourable parameters. After which, an investigation on the effects of different individual data augmentations in Section 5.5, and then finally an investigation on the best combined data augmentations in Section 5.6. Throughout all the investigations, both the final model and the model with the highest validation dice score in the training process are saved and will be evaluated in the test set investigations.

5.1 Baseline Model

The pipeline used for the baseline model and all the future investigations is shown in section 4.4. The baseline U-NET model used was identical to the original [44] except with the addition of batch normalisation layers after every convolution in the encoding path, as presented in [46]. The hyperparameters used for this model are shown in table 4.

Hyperparameter	**Value**	**Reason**
Image Size	256x256	Using an image of this size allows the use of a larger batch size. Also, this is a typical image size to use in this task [46] [55].
Batch Size	16	Between 4 and 32, which are the batch sizes that can be run on the GPU machine without running out of memory.
Loss function	Dice Loss	Used in the following papers [46].
Optimizer	Adam	Widely used in breast cancer segmentation [54] [55] [46].
Learning Rate	$1e^{-4}$	Typical values range between $1e^{-3}$ and $1e^{-5}$, therefore the midpoint was chosen.
Epochs	100	Arbitrarily chosen.

Table 4: Table of hyperparameters used in the baseline model investigations.

From the validation set results shown in figure 17, we see that the model overfits to the training data, as shown by the increasing disparity between the training dice score and the validation dice score as the number of epochs increases. In addition to that, we see that the validation dice score reaches its peak just before epoch 20, after this epoch the validation dice score hovers between 0.55 and 0.6. At the end of the 100

epochs, the training dice score is approximately 35% greater than the validation dice score. Figure 18 showcases a prediction performed by the model. We see that the model underpredicts, as the predicted area of the tumour is much smaller than the ground truth. This is because over half of the filtered dataset are normal mammograms without any tumour/ROI, therefore if the model is trained for too long it learns to only predict true if it is certain that there is a tumour there. This leads to models which have a high precision but a low recall. This is undesirable for a medical segmentation model, as the consequences of an FN are much more drastic than the consequences of an FP.

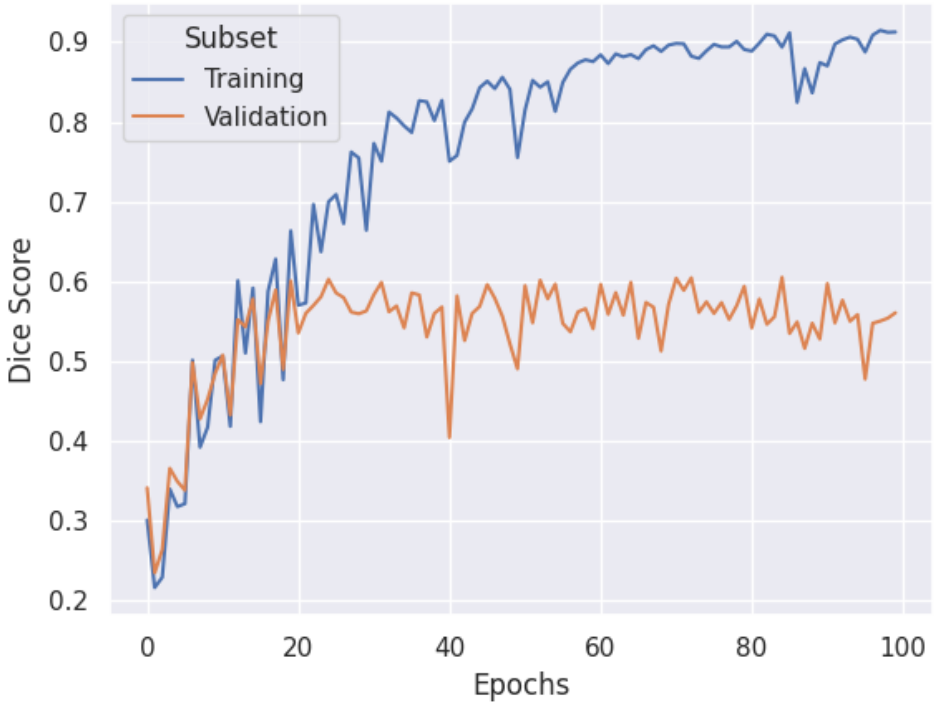
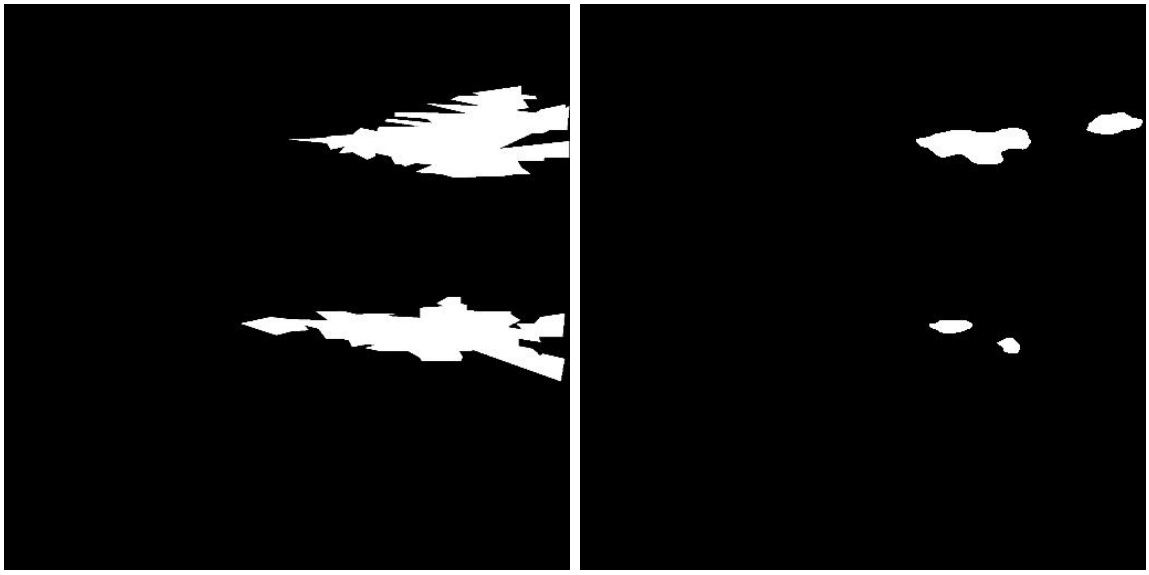


Figure 17: Dice score for the baseline model with no early stopper. Here we see the validation and training dice score drastically diverge from one another after epoch 20.



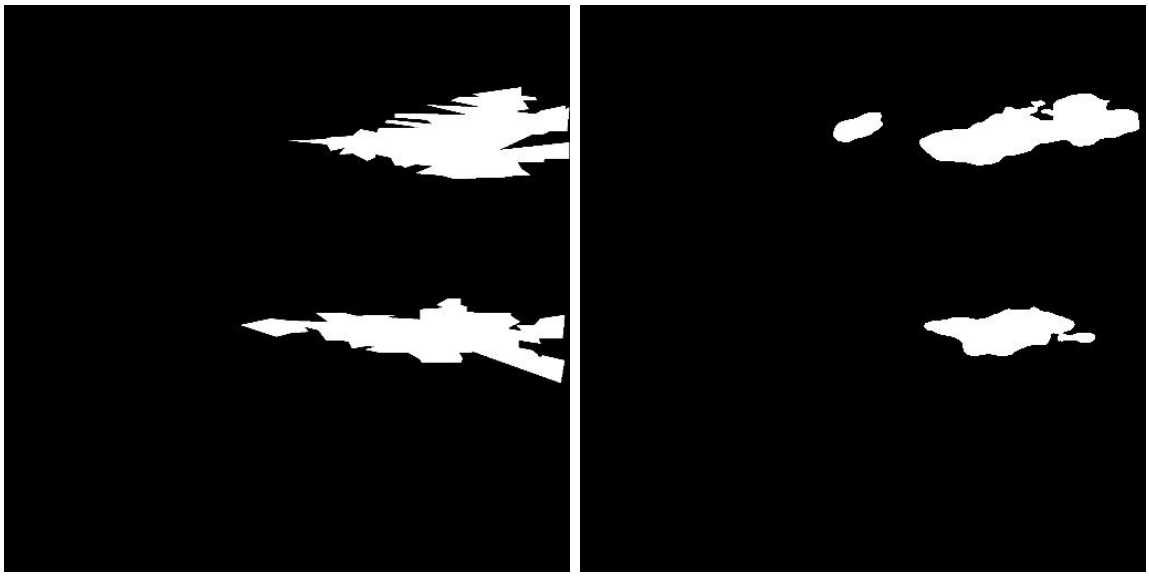
(a) Ground Truth (b) Model Prediction

Figure 18: An example of a prediction of the model with no early stopper. This was done on image *P55_R_CM_CC.jpg* which is in the test set. The model is able to highlight the correct area, but predicts a drastically smaller region than the ground truth.

In order to rectify this issue, an early stopper was implemented, which would halt training if the model was beginning to overfit to the training data. The Δ_{min} parameter is the threshold representing the minimum difference required for the training dice score to be considered to be diverging from the validation dice score, indicating that the model was beginning to overfit, this parameter was set to 0.03. The *patience* parameter specifies the number of consecutive epochs where the difference between the training and validation dice scores exceed the Δ_{min} threshold before the early stopper halts the training process.

The validation set results using the early stopper are shown in figure 19. We see that the early stopper is effective as it halts the training process after 26 epochs, before the model has a chance to overfit to the training images. The validation dice score with the early stopper was 0.580, this was 0.02 better than without the use of an early stopper. In addition to that, the predictions that the model makes are much better, as shown in figure 20.

Figure 19: Dice score for the baseline model with an early stopper. We see the training gets cut off much earlier than the specified 100 epochs of training.



(a) Ground Truth (b) Model Prediction

Figure 20: An example of a prediction of the model with the early stopper. This was done on image $P55\_R\_CM_CC.jpg$ which is in the test set. The prediction area is much closer to the ground truth compared to figure 18.

Early Stopper	Final Training Dice Score	Final Validation Dice Score	Training Time (s)
Yes	0.913	0.561	1839
No	0.673	**0.580**	503

Table 5: Final Results from the baseline model investigation

5.2 Hyperparameter Tuning

This investigation aims to find the best settings for the hyperparameters of the model which maximises the final dice score. A Grid Search was used to find the best combination of hyperparameters. This method was chosen as it is not only easy to implement, but also is able to find the best combination of hyperparameters within the ranges that we have set. However, due to the fact that Grid Search iterates through all possible combinations, we have to be careful that the search space is not too large, and that this investigation can be completed in a sensible amount of time. Due to the power of the machines that the investigations were run on, there were not a lot of possible image size and batch number combinations that could be run without running out of VRAM, drastically reducing the search space we have to investigate over. Furthermore, the maximum number of epochs was reduced from 100 to 75, to further reduce training time.

Another factor that we are investigating is the effect that a decreasing learning rate has on the final result of the model. Varying the learning rate during training is a very popular technique that has shown to minimise the final loss of machine learning tasks [80]. The technique chosen to decay the learning rate is shown in equation 5.

$$lr_{new} = lr_{starting}(1 - \frac{e_{current}}{e_{max}})^2 \tag{5}$$

This is a polynomial decaying learning rate which starts at an initial learning rate, $lr_{starting}$, and slowly decays to a minimum learning rate, lr_{min}. The learning rate for a particular epoch is determined by the fraction of the maximum epochs completed, as shown in 5. This implies that initially, the learning rate decreases rapidly, but as we near the final epoch, the learning rate decreases more gradually, with very small steps taken after each epoch, figure 21 showcases how the learning rate changes over 100 epochs.

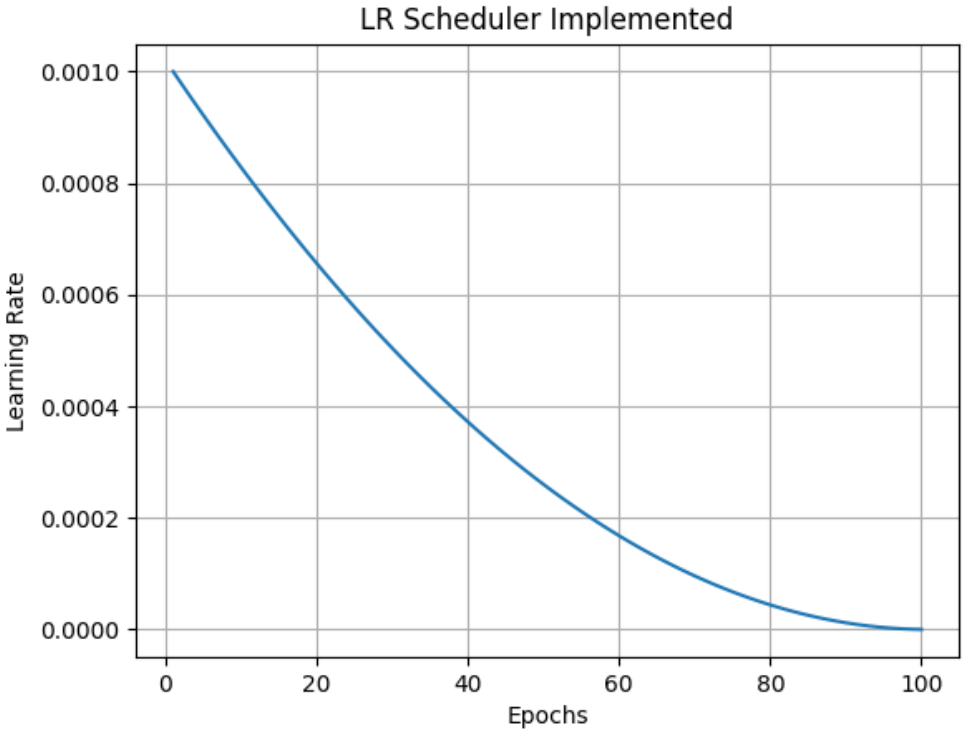


Figure 21: The polynomial learning rate scheduler implemented. The starting learning rate is $1e^{-3}$, the minimum learning rate is $1e^{-6}$, the number of epochs is 100.

Hyperparameter	Range	Justification
Starting LR	$1e^{-3}$, $1e^{-4}$, $1e^{-5}$	Any values larger than $1e^{-3}$ result in the loss values diverging to NaN. These values are also very common in the literature as seen in [55] [45].
Minimum LR	$1e^{-3}$, $1e^{-4}$, $1e^{-5}$, $1e^{-6}$, $1e^{-7}$	The minimum lr has to be equal to or lower than the starting lr.
Image Size	256x256, 512x512	256x256 is very common in the literature as previously mentioned. 512x512 is also tested to check if larger images perform better than smaller images.
Batch Size	4, 8, 16, 32	These have to be a power of 2, we choose to test all possible batch size values that are possible to run with the GPU machine available.

Table 6: The ranges of hyperparameters the grid search is performed on.

Overall, there were 72 different combinations of hyperparameters, which took 22 hours, 34 minutes and 12 seconds to run. Surprisingly, the smaller image sizes on average performed much better than the larger images, as seen in figure 22. This highlights the fact that, image size may play less of a role in the segmentation of breast cancer in this dataset than we previously assumed. Moreover, higher batch sizes were shown to have slightly less variance than the smaller batch sizes, as seen in figure 23. That being said, the best configuration of hyperparameters was one where the batch size was 4. The best configuration is shown in Table 7 and was able to achieve a dice score of 0.627, an improvement of 0.047 over the baseline model on the validation set.

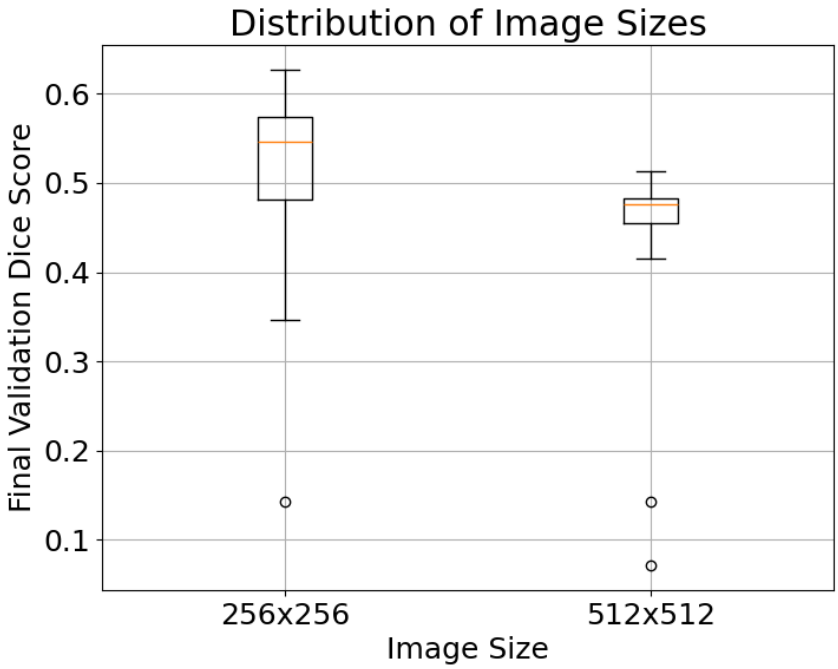


Figure 22: Box plot of the final validation dice scores when grouping together image sizes. Smaller images seem to contribute to a higher dice score, however, have a larger variance.

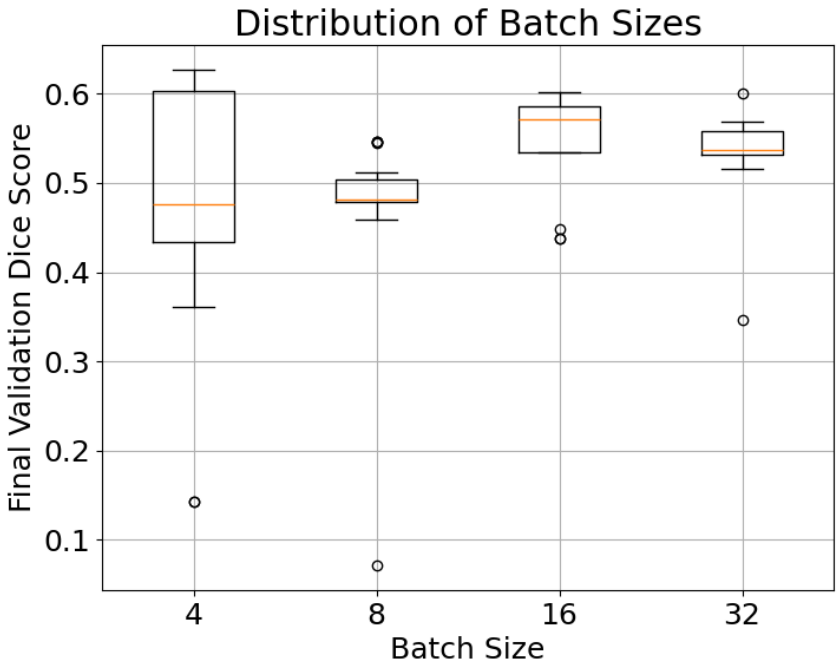


Figure 23: Box plots of the final validation dice scores when grouping together batch sizes.

Starting LR	$1e^{-4}$
Minimum LR	$1e^{-5}$
Batch Size	4
Image Size	256x256
Dice Score	0.627

Table 7: The combination of hyperparameters, which produced the greatest final validation dice score. These hyperparameters were used for the following investigations. The results of all the different hyperparameters tested is shown in Appendix 9.1.

This investigation showcases the importance of the choices of hyperparameters and how hyperparameter tuning is a necessary step for all deep learning investigations.

5.3 Loss Function and Optimiser Function

The combination of loss and optimiser function used is incredibly important in the field of deep learning. Therefore, an investigation was performed to identify the best combination of these functions for our model. The three loss functions investigated were: Dice loss, Dice Binary Cross Entropy (BCE), and Tversky loss. The three optimisers investigated were: Adam, Adamax, and RMSProp.

5.3.1 Dice Loss

The Dice loss is simply defined as being $1 - DS$, where DS is the Dice Score. It is a very simple metric that attempts to maximise the dice score, the primary metric we are using to determine model performance. In addition to this, due to how the dice score is calculated, it is well suited to deal with the problem of the foreground being much smaller than the background [81], and is therefore appropriate for our use.

5.3.2 Dice Binary Cross Entropy Loss

Binary Cross Entropy is a common loss function that is often used in binary classification tasks, and is defined by equation 6.

$$L = -\frac{1}{N}\sum_{n=1}^{N} y_n \; log(\hat{y_n}) + (1 - y_n) \; log(1 - \hat{y_n}) \tag{6}$$

The loss function is widely used in the field of machine learning due to the fact that its derivative is very easy to calculate and that the function is very stable, making it perfect to use alongside gradient descent. This loss function can also be applied to a binary segmentation task, where the loss of a prediction and its ground truth is simply the mean of the pixel-wise BCE losses. Unfortunately, this loss function can perform poorly in unbalanced datasets. Therefore, the Dice BCE loss can be defined by simply adding the Dice loss and BCE loss together, thus leveraging the benefits of both BCE and Dice Loss in one function.

5.3.3 Tversky Loss

Another loss function that will be experimented will be the Tversky loss function. The Tversky loss function is a variant of the dice loss function which attempts to weigh FNs more than the FPs, which is crucial when dealing with imbalanced data [82]. Two parameters α and β are used to weigh the FPs and FNs respectively, as can be seen in equation 7.

$$L = \frac{TP}{TP + \alpha FP + \beta FN} \tag{7}$$

In order to find the best values for α and, β an experiment was run to iterate over 5 different assignments of the parameters, as is done in [82]. The table of results is shown in table 8

Parameters	Final Validation Dice Score
$\alpha = 0.5$, $\beta = 0.5$	0.586
$\alpha = 0.4$, $\beta = 0.6$	0.575
$\alpha = 0.3$, $\beta = 0.7$	0.508
$\alpha = \mathbf{0.2}$, $\beta = \mathbf{0.8}$	**0.600**
$\alpha = 0.1$, $\beta = 0.9$	0.463

Table 8: The different settings of α and β tested for the Tversky loss. The Adam optimiser was used for all the parameters. The parameters tested were the same as in [82]. Values $\alpha = 0.2$ and $\beta = 0.8$ were used for the following investigation.

5.3.4 Experimental Results

Overall, this investigation took 2 hours 39 minutes and 32 seconds to run and showcased that the dice loss function with the Adam optimiser performed the best out of all combinations, see Table 9. This model achieved a validation dice score of 0.627, 16% higher than the worst pairing being the Dice BCE loss function and the Adam optimiser with a dice score of 0.466. From the results, we also see that, on average, the Dice Loss performs much better, with an average dice score of 0.589, than the Dice BCE and Tversky loss functions which had an average dice score of 0.506 and 0.553. This reveals the importance of finding a good loss function that can work well on the dataset being used. Conversely, we see that the choice of optimiser has, on average, less of an effect on the performance with the model with all three average dice scores being closer to one another, 0.564, 0.528, 0.556.

	Dice Loss	Dice BCE Loss	Tversky Loss
Adam	**0.627**	0.466	0.600
Adamax	0.562	0.512	0.510
RMSProp	0.577	0.542	0.550

Table 9: Final Validation Dice scores for all the loss and optimiser function pairings. The Adam optimiser with the dice loss function performed the best. This pairing will be used in the future investigations.

5.4 Seed Search

In machine learning, seeds are used to in the initialisation of weights and to ensure that the experiments are reproducible by the use of seeded random number generators. During gradient descent, the initial model parameters (weights, and biases) are incredibly important, as they can be defined as the starting position of the model in the model parameter 'space'. The goal of gradient descent is to minimise the specified loss function by changing the model's parameters using the gradient of the loss function to traverse this space. Therefore, if we start in a more favourable position, ideally near the

global minimum, we have a higher likelihood of converging more quickly towards the global minimum. Conversely, if the model were to start in a poor position, the model may never reach the global minimum as it gets 'stuck' in a local minimum, figure 24 showcases this in a 2D contour plot.

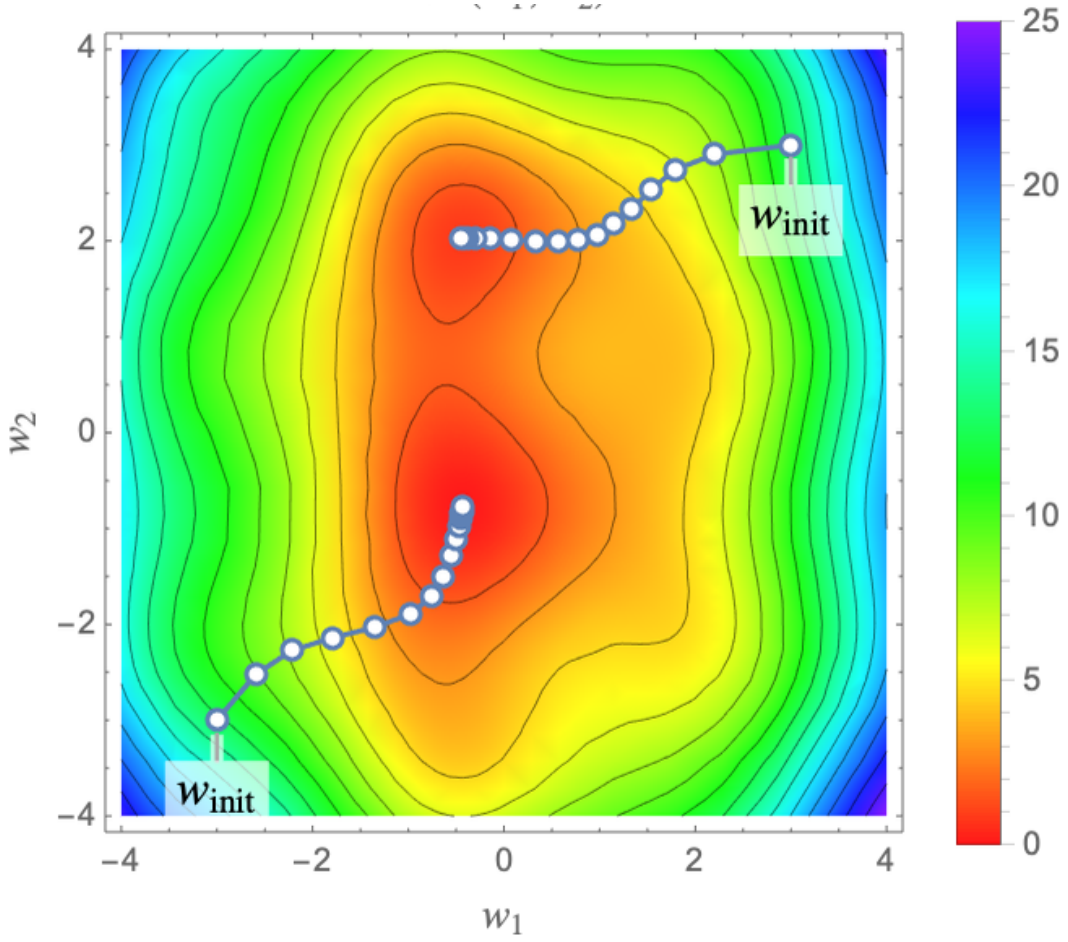


Figure 24: A contour plot showcasing how a different initialisation of parameters, weight 1 and 2 in this case, can cause the model to converge to different minimas within the space. Taken from [83]. Here we see that the minimas represent the values of the two weights that minimise the loss function.

The above example, is a very simple one, however in our case we are dealing with a very high dimensional space which will have many more local minimas where our model can get stuck. Therefore, it is imperative that we attempt to find a seed where our dice score is maximised. Realistically, it would be impossible to iterate through the entire possible set of seeds, therefore the experiment will be conducted on 10 random seeds.

Overall, this investigation took 1 hour and 45 minutes to run. From the investigation results, see Figure 25, we see that the best seed over the 10 that were tested is seed

42 with a final dice score of 0.627. Much like with the hyperparameter tuning and the loss and optimiser function choice, we see that experimenting over different seeds is incredibly important to ensure the best possible performance. The worst seed had a dice score of 0.497, 13% lower than the best seed, this highlights how the local minimums that the model can get 'stuck' in can be significantly greater than the global minimums. Additionally, this disparity between the maximum and minimum dice score also illustrates how bumpy this parameter space is, further reinforcing the importance of performing a seed search in machine learning tasks.

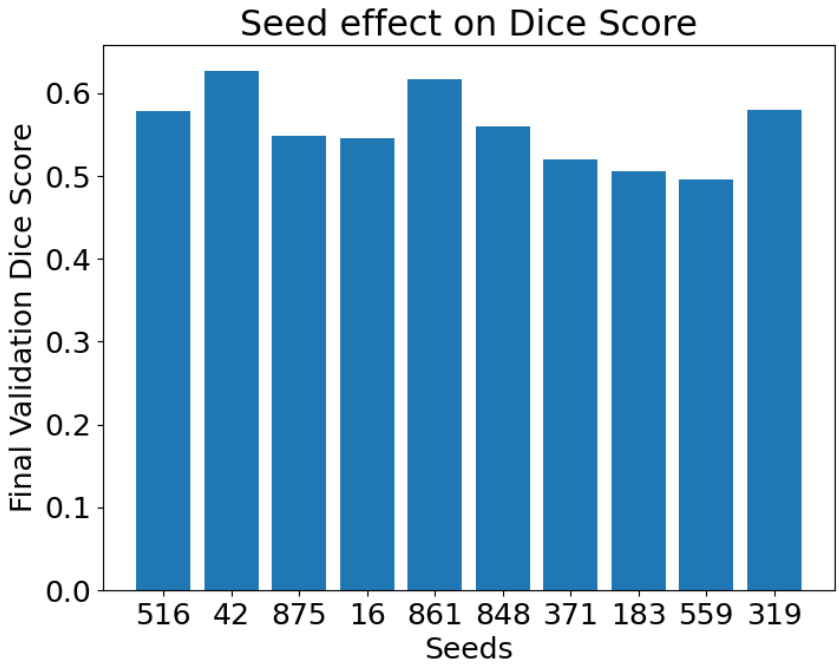


Figure 25: Bar chart of the final dice scores for the validation set for the different seeds.

From all the previous investigations, we see that the best tuned model achieved a dice score of 0.627 with the parameters shown in Table 10. We will refer to this model as the **Tuned Baseline** model from now onwards.

Parameter	Value
Starting LR	$1e^{-4}$
Minimum LR	$1e^{-5}$
Image Size	256x256
Batch Size	4
Loss Function	Dice Loss
Optimiser Function	Adam
Seed	42

Table 10: The final settings for each of the parameters for the tuned model.

5.5 Individual Data Augmentation Techniques

This investigation aims to find the best parameters for each of the data augmentation techniques for our model. It would be impossible to iterate through all possible parameters for each of the data augmentation techniques, therefore the search has been limited to the values shown in table 11.

Data Augmentation	**Values**
Horizontal Flips	On/Off
Vertical Flips	On/Off
Rotations	5°, 10°, 20°, 45°, 60°, 90°
X-Shearing	0.1, 0.25, 0.5, 0.75, 1.0
Y-Shearing	0.1, 0.25, 0.5, 0.75, 1.0
Translation	25, 50, 75, 100 pixels

Table 11: The choices of individual data augmentation parameters to search on. Note that rotation produces 4 rotated versions of the image, if the rotation parameter is r, we produce images that have been rotated by these amounts: $-2r$, $-r$, r, $2r$

We perform a search for the best individual parameters for each of the data augmentations, in order to reduce the search space of the grid search performed in the next investigation. If we were to perform a grid search on all possible combinations of data augmentations and their parameters, 11315 number of experiments would have to be performed. If we assume that each experiment takes 30 minutes, a low estimate since the training times would drastically increase due to the increase in the training set size, this would take 235 days to run.

Overall, this investigation took 9 hours, 10 minutes and 5 seconds to run over all 22 augmentation parameters.

5.5.1 Flipping Images

From the investigations, we see that both type of flipping augmentations reduce the final validation dice score, see Table 12. This is very surprising because these techniques are widespread in the literature. In addition to that, it is surprising that horizontal flips do not increase the performance of the model, as the images added into the dataset are within the domain of images already present. Not only did the model perform worse with the addition of these augmentations, the models took twice as long to run due to doubling the training set size.

Flip Type	Dice Score
Horizontal	**0.603**
Vertical	0.591

Table 12: Final Validation dice score after when adding horizontal and vertical flips separately to the original training set images. Bold values highlight the parameters and augmentations use in the grid search in the following investigation.

5.5.2 Rotation of Images

From the investigations we see that rotation is a very useful augmentation in this context as two of the 6 different rotation angles, $5°, 10°$, performed better than the tuned model without data augmentation, achieving a dice score of 0.643 and 0.650 respectively. In addition to that, we see a trend that the smaller rotations typically outperform the more drastic rotations. This highlights the fact that rotations of a smaller degree, $< 20°$ can lead to a more generalisable model compared to not using any augmentation.

Angle (°)	Dice Score
5	0.643
10	**0.650**
20	0.614
45	0.544
60	0.528
90	0.558

Table 13: Final Validation dice scores after adding all the different rotated images separately to the original training set images. Rotations with parameter 5 and 10 degrees improve the model's performance over not using augmentation techniques. Bold values highlight the parameters and augmentations use in the grid search in the following investigation.

5.5.3 Shearing of Images

From the results of the investigations we see that like the introduction of flipped images, the addition of shearing images provide no increase in the performance of the model, in fact it has led to a decrease in model performance. Surprisingly, there seemed to be no relation between the severity of the shearing and the model performance, as we see that an x shear of 0.25 and 1 lead to a very similar final dice score. Much like with flipping, it seems that the addition of sheared images to a mammogram dataset provides no benefit to the model's performance.

Parameter	Shear Direction	Dice Score
0.1	x	0.557
	y	0.529
0.25	x	0.596
	y	0.541
0.5	x	0.591
	y	0.594
0.75	x	0.530
	y	**0.617**
1.0	**x**	**0.600**
	y	0.553

Table 14: Final Validation dice scores after adding all the different rotated images separately to the original training set images. Bold values highlight the parameters and augmentations use in the grid search in the following investigation.

5.5.4 Translation of Images

Finally, from the results of the investigations 15, the addition of translated images provides no benefit to the performance of the model, much like the addition of flipping and sheared images. However, the results indicate that a greater translation generally provides a better dice score compared to smaller translations. Considering that the images have already been cropped to minimise the background in the image, translation in the x direction will lead to a part of the image being moved outside the 256x256 region of the image. This will lead to some areas marked as being benign/malignant being moved out of the image, increasing the already large imbalance between the number of pixels being classed as normal and the number of pixels being classed as cancerous. Therefore, the model can achieve a similar or greater dice scores by simply under-predicting.

Translation (pixels)	Dice Score
25	0.548
50	**0.603**
75	0.600
100	0.603

Table 15: Final Validation dice scores after adding all the different translated images separately to the original training set images. Bold values highlight the parameters and augmentations use in the grid search in the following investigation.

5.5.5 Overall Effect of Individual Data Augmentations

Overall, we see that a majority of the individual data augmentations result in the model performing worse than the tuned baseline, which did not have any data augmentations. In fact, only rotations of 5 and 10 degrees resulted in a performance increase over the tuned baseline. Furthermore, apart from rotations, there did not appear to be a relationship between how realistic the augmented image was, and its subsequent effect on the performance of the model.

5.6 Combined Data Augmentation Techniques

This investigation aims to find the best combination of the individual data augmentation techniques for the CESM dataset. In this section of the investigation *combination* refers to the aggregation of the different individual data augmented images into one training set. While most of the individual augmentations made the model perform worse, we do not know what effect combining these augmentations may have on the model's performance. The parameters used for each augmentation are the ones that performed the best in the previous investigation. The only exception to this was the decision to exclude the vertical flips on the basis that it performed the worst compared to the other augmentations. This drastically reduced our search space from 57 different combinations to 26.

Out of the 26 different combinations, only 6 achieved a greater dice score than the tuned model without any data augmentations, of which only 2 scored greater than the best individual data augmentation, see Table 16. The observed decrease in average performance with the addition of the individual and combined data augmentations might be attributed to the introduction of excessive variance into the training data. This additional variance in the training data may have made it harder for the model to find meaningful patterns or features present in the original mammogram images, in other words the model overfits to the variance present in the training set, hindering its generalisability. Another observation is that simply combining more data augmented images together is not a 'fix-all' solution to the lack of data. Once again, this is most likely due to the addition of more variance in the training set.

Hflip	Shear X	Shear Y	Translate	Rotate	Dice Score
✓	✓	✓	✗	✗	0.648
✓	✗	✗	✗	✓	0.641
✓	✗	✗	✓	✓	0.652
✗	✗	✓	✗	✓	0.629
✗	✗	✓	✓	✓	0.639
✗	✗	✓	✓	✗	**0.655**

Table 16: The combination of individual augmentations that achieved a dice score greater than the dice score achieved by the tuned model without any data augmentations, which was 0.627. The table containing all the results from this investigation is shown in Appendix 9.1.

The second part of this investigation aims to look at the effect of a different method of combining these augmentations. Instead of aggregating the different individual augmented images together into one set, the augmentations would be applied during the training process of the model. To do this, the rotation set of 10 degrees was chosen as the training set, as this was the best individual augmentation. The Albumentations python library was used to apply the different data augmentations to the training images when they were retrieved at every epoch. The different augmentations had different probabilities of being applied to the image and the mask. This was done so that the images in the training set would be different for each epoch, hopefully decreasing the effect of overfitting and increasing generalisability. The transformations applied to the images was a composition of: an x and y shear with a 0.2 probability, a vertical flip with a 0.5 probability, and a horizontal flip with a 0.5 probability.

This experiment had a final validation dice score of 0.694, higher than any investigation before this. However, from Figure 26 we see that the validation dice score is consistently higher than the training dice score, even at the final epoch. This indicates the early stopper might be stopping the training prematurely.

Figure 26: The training and validation dice score at each epoch when composing different augmentations onto one image using the Albumentations library.

The experiment was repeated with the removal of the early stopper, a final dice score of 0.632 was achieved, much lower than the previous result. We see from figure 27 that the model exhibits the expected behaviour of the training dice score being slightly above the validation dice score. While the final validation dice score may be lower than the previous experiment, the added training epochs may allow better generalisation in the test set.

Figure 27: A repeat of the experiment shown in Figure 26 without an early stopper, letting the model train for the full 75 epochs.

6 Test Set Results

This section aims to evaluate the best models for each individual investigations in Section 5. The models will be tested on the unseen test subset of the dataset which contains a total of 114 images, 65 normal images, 9 benign images, and 40 malignant images. No data augmentation techniques were performed on these images. The test set results serve as the true performance of the model if it were to be used in real-world applications.

In addition to the dice score metric, the precision, recall, and accuracy of the model's performance on the test set will also be recorded. During the model's training, the model's state was saved after the final epoch as well as when the validation dice score was its highest, both of these models will be evaluated on the test set for each of the investigations.

The results of the baseline model and the tuned model with no data augmentations are shown in Table 17. We see that the model that performs the best is the tuned model at the point of its maximum dice score during training. The tuned model performs 5% better than the baseline model on the test set, achieving dice scores of 0.526 and 0.476 respectively.

Investigation	Model	Dice Score	Precision	Recall	Accuracy
Baseline	Last Epoch	0.476	0.425	0.648	0.959
	Max Dice	0.446	0.471	0.633	0.964
Hyperparameter	Last Epoch	0.509	0.505	0.566	0.967
	Max Dice	**0.526**	0.641	0.499	0.974

Table 17: The results of the investigations used to tune the model. The seed is set to 42, as it was the best performing seed during the validation set investigations. We do not showcase the results for the loss and optimiser function and seed search models due to the fact that they are the same as the hyperparameter tuned model.

The results of the individual data augmentations are shown in Table 18. Here we see a similar trend to the results in the validation set, where rotation is the only augmentation which improves the performance of the model. The rotation augmentation model achieves a maximum dice score of 0.550. Unlike the validation set, the worst performing augmentation was translation - by a noticeable amount. This can be attributed to the fact that the images have already been cropped to the breast region, thereby making this augmentation at best redundant and at worst disruptive to the model's ability to generalise.

Augmentation	Model	Dice Score	Precision	Recall	Accuracy
Horizontal Flips	Last Epoch	0.497	0.487	0.580	0.966
	Max Dice	0.504	0.560	0.524	0.971
Vertical Flips	Last Epoch	0.506	0.638	0.511	0.974
	Max Dice	0.504	0.614	0.552	0.974
Rotations	Last Epoch	0.537	0.619	0.540	0.974
	Max Dice	**0.550**	0.591	0.555	0.973
X Shearing	Last Epoch	0.497	0.617	0.462	0.973
	Max Dice	0.457	0.63	0.492	0.973
Y Shearing	Last Epoch	0.498	0.546	0.557	0.970
	Max Dice	0.521	0.579	0.543	0.972
Translation	Last Epoch	0.458	0.625	0.510	0.974
	Max Dice	0.472	0.652	0.492	0.974

Table 18: The results of the effect of the individual augmentations on the test set. The parameters used for each of the augmentations are the ones that performed the best on the validation set.

The results of the combined data augmentations are shown in Table 19. Much like the validation set, the combination of a y shear of 0.75 and a translation of 50 pixels performed the best, obtaining a dice score of 0.566. However, unlike the validation set, a half of the combinations perform worse than the tuned model with no augmentations.

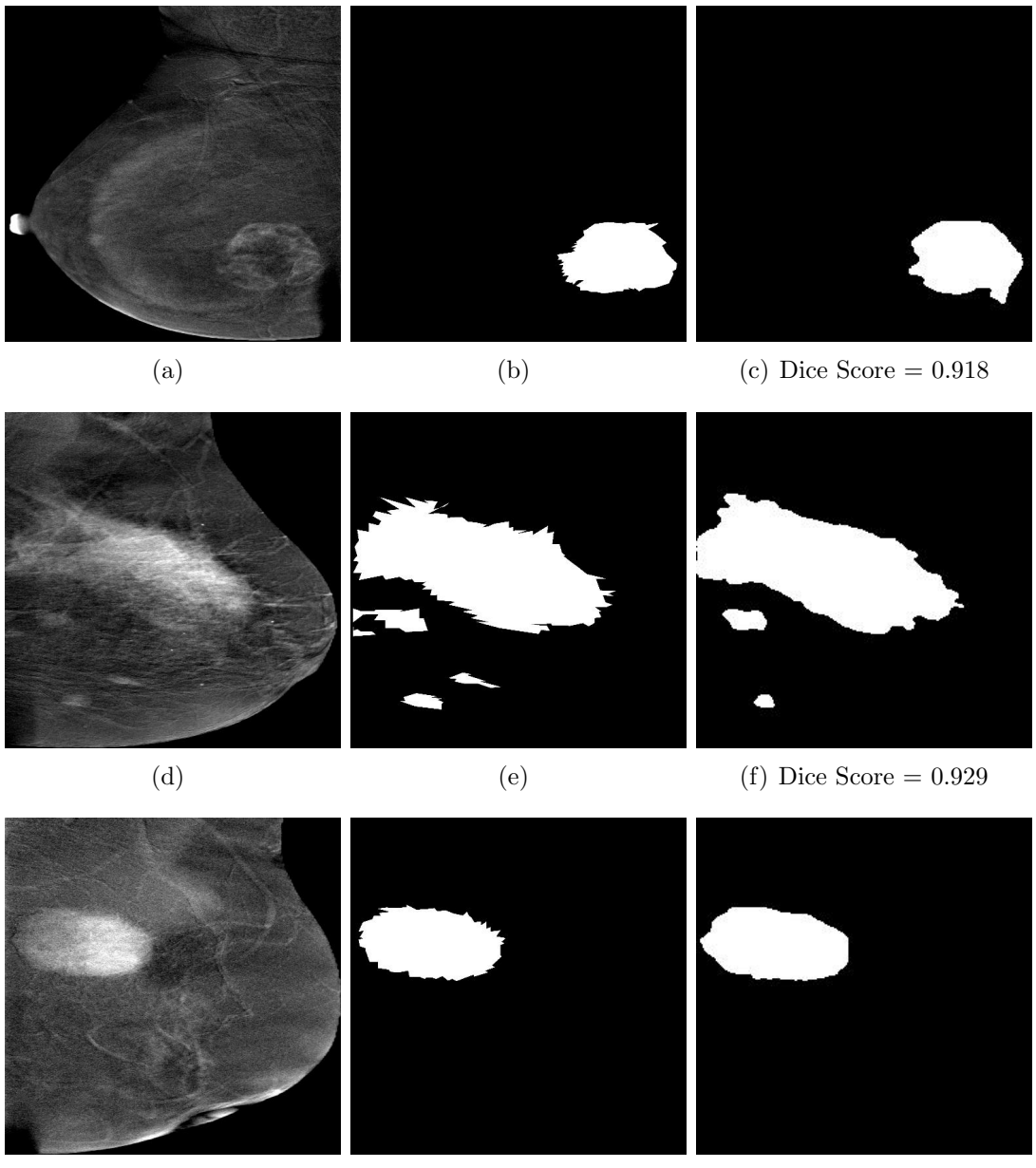
Surprisingly, the use of the Albumentation augmentations on the rotation set decreased performance compared to just using the rotation set alone. This is surprising as this library is often used to add variation in the training set data, improving the model's performance on the test set. However, it is possible that there are too few images in the training set to take advantage of this fact. Alternatively, perhaps the composition of different augmentations resulted in images that differed too much from the original training images.

Augmentation Grouping	Model	Dice Score	Precision	Recall	Accuracy
HFlip + X Shear + Y Shear	Last Epoch	0.523	0.579	0.505	0.971
	Max Dice	0.499	0.569	0.504	0.971
HFlip + Rotation	Last Epoch	0.529	0.573	0.538	0.971
	Max Dice	0.529	0.573	0.538	0.971
Hflip + Translation + Rotation	Last Epoch	0.472	0.709	0.465	0.976
	Max Dice	0.554	0.581	0.594	0.972
Y Shear + Rotation	Last Epoch	0.476	0.603	0.494	0.972
	Max Dice	0.492	0.580	0.543	0.972
Y Shear + Translate + Rotation	Last Epoch	0.540	0.682	0.489	0.976
	Max Dice	0.500	0.524	0.597	0.969
Y Shear + Translate	Last Epoch	**0.566**	0.650	0.514	0.975
	Max Dice	0.566	0.650	0.514	0.975
Rotation + Albumentation + early stopper	Last Epoch	0.494	0.510	0.557	0.968
	Max Dice	0.494	0.510	0.557	0.968
Rotation + Albumentation + no early stopper	Last Epoch	0.527	0.521	0.584	0.968
	Max Dice	0.494	0.510	0.557	0.968

Table 19: The results of the effect of the individual augmentations on the test set. The parameters used for each of the augmentations are the ones that performed the best on the validation set.

The model that performed the best on the hidden test set was the tuned model with the following combined augmentation: a y shear of 0.75 and an x translation of 50 pixels. This model achieved a dice score of 0.566. Figure 28 showcases some predictions the best model made on a few of the images of the test set. The model seems to perform the best on images where the suspicious mass is the highest contrast region of the image. In addition to that, many of the predictions that perform well, dice score > 0.9, have ground truth masks with very large regions segmented. Figure 29 showcase poor predictions made by the model. We see that the model predictions are noticeably worse for masses which are relatively smaller than the ones shown in Figure 28. This highlights a potential flaw in the dice score, where smaller mistakes are more heavily

penalised on masks with smaller regions. This is especially noticeable on images which are a part of the normal class, where the ground truth mask is empty, where a single pixel prediction will result in a dice score of 0. This effect could have been mitigated by a post-processing step that would remove very small segmented regions from the predictions.

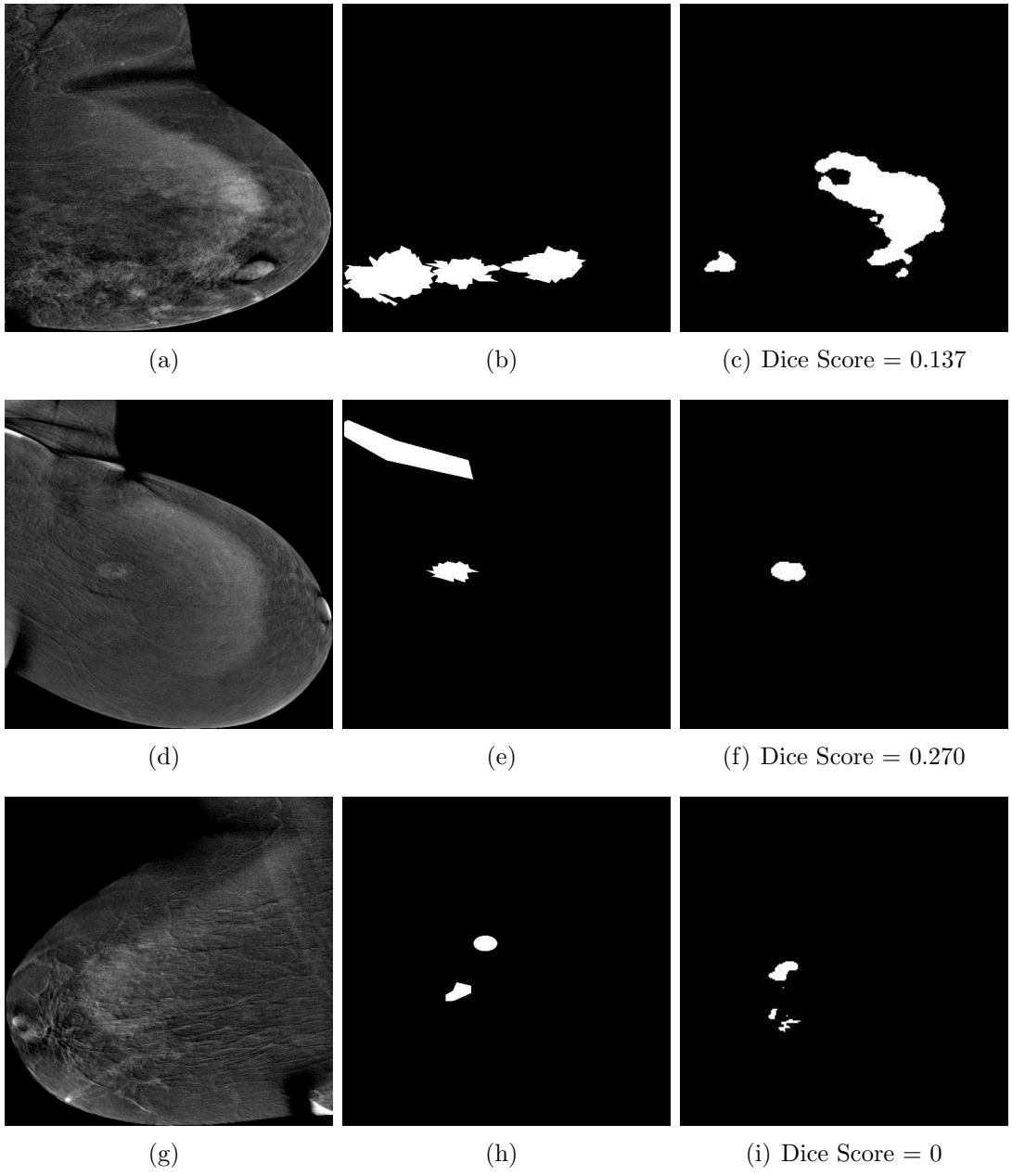


(a) (b) (c) Dice Score = 0.918

(d) (e) (f) Dice Score = 0.929

(g) (h) (i) Dice Score = 0.956

Figure 28: Good predictions made by the best performing model on some of the images in the test set. Column 1 showcases the original mammograms, column 2 showcases the ground truth segmentations, and column 3 showcases the model predictions.



(a) (b) (c) Dice Score = 0.137

(d) (e) (f) Dice Score = 0.270

(g) (h) (i) Dice Score = 0

Figure 29: Bad predictions made by the best performing model on some of the images in the test set. Column 1 showcases the original mammograms, column 2 showcases the ground truth segmentations, and column 3 showcases the model predictions.

Overall across all investigations, the performance of the models on the test set is significantly lower than that on the validation set. This signifies the fact that the images present in the training and validation set are not representative of the entire dataset as a whole. Furthermore, this could highlight a possible class imbalance in the training-

validation-test split. When splitting the data, benign and malignant instances were considered as one class due to the relatively low number of benign cases present in the dataset.

Another surprising result of the models on the test set is the relatively small benefit, if any, that data augmentation seem to have on this particular dataset. This unexpected outcome could suggest that the inherent characteristics of the dataset may have rendered data augmentations less effective than we otherwise anticipated.

The results on the test set also highlight how the early stopper may have possibly been too strict and halted training prematurely. This then resulted in some of the models slightly underfitting to the data, leading to lower performance. Alternatively, this discrepancy could be a result of the dataset itself. For example, the labels given to each of the images could have been not entirely accurate, thereby making the removal of redundant images based on their labels ineffective. This would have kept a large variance between the photos, as images which should of been removed were kept.

7 Discussion

7.1 Critical Appraisal

In this dissertation, a breast segmentation model was implemented and refined to segment breast cancer from CESM images from a new publicly available dataset [1]. This dataset was chosen as it is the first publicly available dataset that was solely composed of CESM images. Furthermore, the dataset has never been used for a segmentation task in the literature. Therefore, this dissertation provides a starting point for the literature surrounding the segmentation of CESM images, that researchers can build on. A primary task of this dissertation was to evaluate the effectiveness of different affine data augmentation techniques on the segmentation performance of a UNET model on the dataset. This specific type of data augmentation was focused on due to how easy it is to generate the new images, especially when compared to more complicated forms of data augmentation such as GANs and elastic deformations. This deliberate choice broadens the applicability of this dissertation's findings, as it enables others performing a similar task to easily implement similar augmentations without having to run a large grid search over a certain parameter space.

In contrast to some prior approaches, no pre-processing techniques were performed on the data in this dissertation. For instance, Zeiser et al. utilised both regions of interests (ROIs) and the original mammograms [45]. A total of 9 ROIs were extracted from each of the mammograms from the DDSM dataset. In cases where a mammogram contained a mass, a 256x256 window was positioned to encompass the mass, after which a combination of 3 different data augmentations including flipping, as well as two different zoom levels, were used to generate the 9 ROIs. In the absence of a discernible mass, 9 random ROIs were extracted from the image. This approach not only maintains the balanced distribution between images depicting masses and those without, but also addresses the issue of the cancerous region being significantly smaller than the 'normal' region.

Considering the prolonged use of datasets such as DDSM and INbreast within the literature, a prevalent and crucial pre-processing step employed in the majority of papers is Contrast Limited Adaptive Histogram Equalisation (CLAHE) [84] [45] [46] [54]. This technique is often used in images with low contrast as a means of enhancing the contrast of an image, without introducing visual artefacts, which is common when using Adaptive Histogram Equalisation (AHE). This enhancement enables any suspicious mass to be more pronounced in the mammogram, enabling a more accurate segmentation diagnosis. While CLAHE may not provide any improvement when used on CESM images, as the imaging modality inherently increases contrast through the use of an injected iodinated contrast agent, exploring other forms of image enhancing pre-processing methods might yield significant performance gains.

Another pre-processing step that has been investigated in the literature is the removal of the pectoral muscle from mammograms. As highlighted in the literature review,

the pectoral muscle can act as potential sources of noise, due to the similar densities between the muscle and potential masses. While there are a plethora of papers introducing new and novel techniques in the segmentation of the pectoral muscle, I could not find any papers that evaluated the effect of this technique in context of segmentation performance of a deep learning model. This gap highlights another potential investigation that could have been performed in the dissertation, offering not only a potential performance gain, but also valuable insight on the impact of this pre-processing step in the context of breast cancer segmentation using deep learning.

Alternatively, another technique that is often seen in the literature is the combination of images from different datasets. For example, Abdelhafiz et al. combined images from 4 different datasets: CBIS-DDSM, INbreast, UCHCDM, and BCDR-01, to create a new dataset containing 2066 mammograms [46]. While only images containing mass segmentations were collated, the same technique could have been used in order to create a balanced dataset of normal, benign, and malignant images. It is unknown whether the combination of CESM and CLAHE DM images would prove fruitful; however, if investigation into this were undertaken, it could yield valuable insights into the potential synergies between the two different imaging modalities. This would allow the addition of more benign and malignant examples to the dataset, potentially reducing the poor recall scores we see the model have on the test set.

Additionally, another facet that could have been explored in the dissertation could have been the use of different configurations of the U-NET model. For instance, Li et al. extended the baseline U-NET by introducing attention gates (AGs) and dense connections [85]. The AGs prioritize relevant features while suppressing irrelevant details of an image, which has been shown to improve performance over the original U-NET [68]. The function of AGs was inspired by our innate ability to perform the same task, which was an evolutionary trait to quickly identify and focus on critical information in a scene. Dense connections, on the other hand, promote feature reuse by allowing each layer in the encoding path to receive inputs from all preceding layers [68]. Huang et al. showcased how these dense connections can achieve higher segmentation scores on small objects, making this a suitable addition in the context of breast cancer imaging [86]. With these additions, the model had a 7% increased dice score compared to the baseline U-NET model on the same dataset. Unfortunately, due to the time constraint of a bachelor's dissertation, it was decided that exploring different models, or possibly even creating a new model, was out of the scope of the project.

Alternatively, instead of manually implementing and comparing different models or configurations of the U-NET, pre-trained models could have been utilised with the Segmentation Model PyTorch library [87]. This would have allowed very quick implementation of different models and configurations that were identified in the literature review. In addition to that, all the models were pre-trained on the ImageNet dataset, meaning that the effect of transfer learning could have also been investigated. This was not performed because, this library was found too close to the submission deadline,

making it infeasible to re-run all the experiments again, while also adding additional investigations.

7.2 Limitations

The main limitation of the dissertation was the time constraints of a bachelor dissertation, this decreased the number of investigations I ran on the dataset. For example, a detailed investigation on the different possible pre-processing techniques could have been applied to the dataset. Techniques such as CLAHE and noise filtering, which are popular when utilising the DDSM dataset, could have been added to the pipeline. Unfortunately, implementing these techniques and then investigating their effects would have been too time-consuming and are therefore out of the scope of this dissertation.

Another limitation was the limitation of computing power of the University GPU machines, which required the images to be resized to 256x256 or 512x512, in order to be used with a batch size greater than 1. The original images of the dataset were cropped to only include the breast region, however a majority of the images were still much larger than 512x512.

7.3 Future Work

In future work, there are many avenues that can be explored in the segmentation of breast cancer using the CDD-CESM dataset. Firstly, given that this dissertation marks the first utilisation of the dataset in a segmentation context, there is an opportunity to explore different pre-processing methods that can be implemented to get the most out of the images in the dataset. Future work could endeavour to explore the potential impact of various pre-processing methods utilized in other datasets, such as DDSM and INbreast, on segmentation performance.

Another potential avenue could involve comparing DM and CESM images. This comparison is especially relevant as the CESM dataset comprises both types of images. Additionally, given the fact that the CESM dataset is relatively small, an investigation of the potential benefits of combining CESM and DM images could prove insightful, as it may allow us to leverage the benefits of the two different image modalities, enhancing the generasability of a model.

Finally, an obvious avenue for future work involves exploring more sophisticated data augmentation techniques. While this dissertation focused on basic affine transformations, such as rotation, translation, scaling, and shearing, there are a plethora of advanced data augmentation techniques that can be explored, as mentioned in Section 2.

7.4 Evaluation

This dissertation met all objectives set out in Section 1.1, see Table 20 for the evidence of completion for each objective.

While the results of the model on the test set fail to compete with the performance seen in the literature, the comprehensive investigations into data augmentations and the exploration of CESM images conducted in this dissertation represent a novel contribution to the literature. Firstly, from the hyperparameter tuning investigations, we see that smaller images (256x256 pixels), consistently outperformed larger images (512x512 pixels). This suggests that downsampling the images in the dataset not only provides a decrease in training and evaluation time, but also a segmentation performance increase. We also see the importance of hyperparameter tuning of segmentation models, as a simple grid search over the hyperparameters increased the model's dice score by 5% on the test set. Additionally, the investigation into individual data augmentations revealed that while augmentations such as rotations of small angles, of 5° and 10°, proved beneficial, simply applying data augmentations did not guarantee improved performance. Notably, augmentations which produced images that differed too much from the original images, such as rotations of 90° and vertical flips, were found to noticeably decrease model performance. We hypothesise that this is a result of the added variation in the training data, which hindered the model's ability to discern relevant features, ultimately leading to decreased segmentation accuracy. Moreover, the combined data augmentation investigations showcase similar results, with only 6 of the 26 combinations performing better than the baseline model on the validation set. This suggests that the addition of the different augmented images to the training set creates too much variation. Overall, the investigations into data augmentations showcase the importance of identifying the optimal set of augmentations tailored for the dataset that one is working on.

The code written for this project is available on my GitHub profile: https://github.com/seanhacks/dissertation.

Objective	Evidence of Completion
Primary: Literature Review	Section 2 showcases a thorough review of the literature surrounding the segmentation of breast cancer using: classical, machine learning, and deep learning methods. In addition to that, topics such as data augmentation, transfer learning (and subsequently image size), and breast cancer datasets were also reviewed.
Primary: Baseline Model and Results	A baseline model and pipeline were implemented using pytorch. The results of this are shown in Section 5.1.
Primary: Investigation of individual data augmentations	The effect of different individual affine data augmentations were investigated in Section 5.5 using a grid search.
Secondary: Investigation of combined data augmentations	The effect of different combinations of the data augmentations were investigated in section 5.6.
Secondary: Fine-tuning of the model	Various hyperparameters of the model as well as the initial seed and the loss, and optimiser functions were optimised using a grid-search. Sections 5.2, 5.3, 5.4 present the results of the tuning.

Table 20: Comparing the work done in this dissertation compared to the initial set of objectives.

8 Conclusion

The aim of this dissertation was to investigate the performance of a U-NET model on the newly released CDD-CESM dataset. A primary investigation of this dissertation was to investigate how the addition of affine data augmentations may increase the segmentation performance on this dataset. In the future, the results from this dissertation can be used to further improve the segmentation performance of CNN models, allowing them to be used as double reading systems. This will decrease the load on radiologists by allowing the CNN to be used as a second opinion instead of other radiologists.

A U-NET model [44] was used for the dissertation with the addition of batch normalisation layers as seen in [46]. First a baseline model and pipeline was constructed to evaluate the model's performance on the dataset without any fine-tuning. This model served as a good reference point for the effectiveness of the fine-tuning and augmentation investigations. The next investigations focused on finding the best set of hyperparameters on the model, this included: image size, starting and minimum learning rate, batch size, loss and optimiser function, and the starting seed. One of the key findings was the fact that the smaller images (256x256), on average resulted in a dice score 7% higher than that of the larger images (512x512). In addition to that, a slight decaying learning rate from $1e^{-4}$ to $1e^{-5}$ over 75 epochs, was found to result in a slightly higher dice score. The next investigation looked at the effect of different individual data augmentations: translation, vertical and horizontal flipping, x and y shearing, and rotation. From the results, we see that only a rotation of 10° resulted in an increase in the model's dice score on both the test and validation set. A key finding from this investigation is that the other augmentations may provide too much variance to the dataset for the dataset used [1]. The final investigation looked at the effectiveness of combining different data augmented images into one set. Much like the previous investigation, only two combinations resulted in a better dice score on the test set: Y shear of 0.75 + translation of 50 pixels, and horizontal flips + translation of 50 pixels + rotation of 10°. The combination that performed the best on the validation set was the use of the rotation set and the albumentation library to apply different compositions of data augmentations with a certain probability, this achieved a dice score of 0.694. The combination that performed the best on the test set was the combination of y sheared and translated images which achieved a dice score of 0.566.

The investigations performed in this dissertation showcase the use of the CDD-CESM dataset in a segmentation task. The primary contribution of this dissertation are the results of a series of thorough data augmentation investigations on this dataset. The results of this dissertation will hopefully aid in the creation of more accurate CAD of breast cancer, which can be used as a double reading system. This would alleviate the burden on radiologists while enhancing productivity, thereby facilitating quicker delivery of scan results to patients.

References

[1] Rana Khaled, Maha Helal, Omar Alfarghaly, Omnia Mokhtar, Abeer Elkorany, Hebatalla El Kassas, and Aly Fahmy. Categorized contrast enhanced mammography dataset for diagnostic and artificial intelligence research. *Scientific Data*, 9(1), Mar 2022.

[2] American Cancer Society. Key statistics for breast cancer. https://www.cancer.org/cancer/types/breast-cancer/about/how-common-is-breast-cancer.html, 2023. Accessed: 18-10-2023.

[3] Rebecca L. Siegel, Kimberly D. Miller, Hannah E. Fuchs, and Ahmedin Jemal. Cancer statistics, 2022. *CA: A Cancer Journal for Clinicians*, 72(1):7–33, Jan 2022.

[4] Cancer Research UK. Breast screening. https://www.cancerresearchuk.org/about-cancer/breast-cancer/getting-diagnosed/screening-breast, 2023. Accessed: 10-11-2023.

[5] Joann G. Elmore. Screening for breast cancer. *JAMA*, 293(10):1245, Mar 2005.

[6] R. Schulz-Wendtland, M. Fuchsjäger, T. Wacker, and K.-P. Hermann. Digital mammography: An update. *European Journal of Radiology*, 72(2):258–265, Nov 2009.

[7] Margaret T. Mandelson, Nina Oestreicher, Peggy L. Porter, Donna White, Charles A. Finder, Stephen H. Taplin, and Emily White. Breast density as a predictor of mammographic detection: Comparison of interval- and screen-detected cancers. *JNCI: Journal of the National Cancer Institute*, 92(13):1081–1087, 07 2000.

[8] Eva M. Fallenberg, Florian F. Schmitzberger, Heba Amer, Barbara Ingold-Heppner, Corinne Balleyguier, Felix Diekmann, Florian Engelken, Ritse M. Mann, Diane M. Renz, Ulrich Bick, and et al. Contrast-enhanced spectral mammography vs. mammography and mri – clinical performance in a multi-reader evaluation. *European Radiology*, 27(7):2752–2764, Nov 2016.

[9] J.J. James and S.L. Tennant. Contrast-enhanced spectral mammography (cesm). *Clinical Radiology*, 73(8):715–723, 2018.

[10] Elzbieta Luczyńska, Sylwia Heinze-Paluchowska, Sonia Dyczek, Pawel Blecharz, Janusz Rys, and Marian Reinfuss. Contrast-enhanced spectral mammography: Comparison with conventional mammography and histopathology in 152 women. *Korean Journal of Radiology*, 15(6):689–696, Nov 2014.

[11] U. C. Lalji, C. R. Jeukens, I. Houben, P. J. Nelemans, R. E. van Engen, E. van Wylick, R. G. Beets-Tan, J. E. Wildberger, L. E. Paulis, and M. B. Lobbes. Evaluation of low-energy contrast-enhanced spectral mammography images by com-

paring them to full-field digital mammography using euref image quality criteria. *European Radiology*, 25(10):2813–2820, Mar 2015.

[12] Breastcancer.org. Learn why breast density matters. https://www.breastcancer.org/risk/risk-factors/dense-breasts, 2023. Accessed: 27-02-2024.

[13] J.J. James and S.L. Tennant. Contrast-enhanced spectral mammography (cesm). *Clinical Radiology*, 73(8):715–723, 2018.

[14] Phillip H. Meyers, Charles M. Nice, Hal C. Becker, Wilson J. Nettleton, James W. Sweeney, and George R. Meckstroth. Automated computer analysis of radiographic images. *Radiology*, 83(6):1029–1034, Dec 1964.

[15] Fred Winsberg, Milton Elkin, Josiah Macy, Victoria Bordaz, and William Weymouth. Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis. *Radiology*, 89(2):211–215, Aug 1967.

[16] Kunio Doi. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, 31(4–5):198–211, 2007.

[17] Maryellen L. Giger, Heang-Ping Chan, and John Boone. Anniversary paper: History and status of cad and quantitative image analysis: The role of medical physics and aapm. *Medical Physics*, 35(12):5799–5820, Nov 2008.

[18] C. Dromain, B. Boyer, R. Ferré, S. Canale, S. Delaloge, and C. Balleyguier. Computed-aided diagnosis (cad) in the detection of breast cancer. *European Journal of Radiology*, 82(3):417–423, 2013. Breast Imaging.

[19] Håkan Geijer and Mats Geijer. Added value of double reading in diagnostic radiology,a systematic review. *Insights into Imaging*, 9(3):287–301, Mar 2018.

[20] The Royal College of Radiologists. New rcr census shows the nhs needs nearly 2,000 more radiologists. https://www.rcr.ac.uk/posts/new-rcr-census-shows-nhs-needs-nearly-2000-more-radiologists. Accessed: 10-11-2023.

[21] Tolga Berber, Adil Alpkocak, Pinar Balci, and Oguz Dicle. Breast mass contour segmentation algorithm in digital mammograms. *Computer Methods and Programs in Biomedicine*, 110(2):150–159, 2013.

[22] R. Adams and L. Bischof. Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(6):641–647, 1994.

[23] Seyedali Mirjalili. Dragonfly algorithm: A new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems. *Neural Computing and Applications*, 27(4):1053–1073, May 2015.

[24] S. Punitha, A. Amuthan, and K. Suresh Joseph. Benign and malignant breast cancer segmentation using optimized region growing technique. *Future Computing and Informatics Journal*, 3(2):348–358, 2018.

[25] L. Shafarenko, M. Petrou, and J. Kittler. Automatic watershed segmentation of randomly textured color images. *IEEE Transactions on Image Processing*, 6(11):1530–1544, 1997.

[26] Marcin Ciecholewski. Microcalcification segmentation from mammograms: A morphological approach. *Journal of Digital Imaging*, 30(2):172–184, Nov 2016.

[27] G. Fotos, A. Campbell, P. Murray, and E. Yakushina. Deep learning enhanced watershed for microstructural analysis using a boundary class semantic segmentation. *Journal of Materials Science*, 58(36):14390–14410, Sep 2023.

[28] K. Santle Camilus, V. K. Govindan, and P. S. Sathidevi. Computer-aided identification of the pectoral muscle in digitized mammograms. *Journal of Digital Imaging*, 23(5):562–580, Oct 2009.

[29] Chen-Chung Liu, Chung-Yen Tsai, Jui Liu, Chun-Yuan Yu, and Shyr-Shen Yu. A pectoral muscle segmentation algorithm for digital mammograms using otsu thresholding and multiple regression analysis. *Computers & Mathematics with Applications*, 64(5):1100–1107, 2012. Advanced Technologies in Computer, Consumer and Control.

[30] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

[31] Xiangyang Xu, Shengzhou Xu, Lianghai Jin, and Enmin Song. Characteristic analysis of otsu threshold and its applications. *Pattern Recognition Letters*, 32(7):956–961, 2011.

[32] Xiang Yu, Shui-Hua Wang, Juan Manuel Górriz, Xian-Wei Jiang, David S. Guttery, and Yu-Dong Zhang. Pemnet for pectoral muscle segmentation. *Biology*, 11(1):134, Jan 2022.

[33] Yihong Gong and Wei Xu. *Machine learning techniques for multimedia*. Cognitive Technologies. Springer US, 1 edition, 2008.

[34] Shubham Sharma, Archit Aggarwal, and Tanupriya Choudhury. Breast cancer detection using machine learning algorithms. In *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, pages 114–118, 2018.

[35] J. C. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57, 1973.

[36] Moustapha Mohamed Saleck, Abdelmajide ElMoutaouakkil, and Mohammed Mouçouf. Tumor detection in mammography images using fuzzy c-means and glcm

texture features. In *2017 14th International Conference on Computer Graphics, Imaging and Visualization*, pages 122–125, 2017.

[37] Wenbin Yue, Zidong Wang, Hongwei Chen, Annette Payne, and Xiaohui Liu. Machine learning with applications in breast cancer diagnosis and prognosis. *Designs*, 2(2):13, May 2018.

[38] Adrian B. Levine, Colin Schlosser, Jasleen Grewal, Robin Coope, Steve J.M. Jones, and Stephen Yip. Rise of the machines: Advances in deep learning for cancer diagnosis. *Trends in Cancer*, 5(3):157–169, Mar 2019.

[39] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, and et al. The cancer imaging archive (tcia): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057, Jul 2013.

[40] Hang Yu, Laurence T. Yang, Qingchen Zhang, David Armstrong, and M. Jamal Deen. Convolutional neural networks for medical image analysis: State-of-the-art, comparisons, improvement and perspectives. *Neurocomputing*, 444:92–110, 2021.

[41] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks, 2015.

[42] Damian Podareanu, Valeriu Codreanu, Sandra Aigner, Caspar Leeuwen, and Volker Weinberg. Best practice guide - deep learning. 02 2019.

[43] Yani Muhamad, Irawan S, and Setianingsih Casi. Application of transfer learning using convolutional neural network method for early detection of terry's nail. *Journal of Physics: Conference Series*, 1201:012052, 05 2019.

[44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer International Publishing, 2015.

[45] Felipe André Zeiser, Cristiano André da Costa, Tiago Zonta, Nuno M. Marques, Adriana Vial Roehe, Marcelo Moreno, and Rodrigo da Rosa Righi. Segmentation of masses on mammograms using data augmentation and deep learning. *Journal of Digital Imaging*, 33(4):858–868, Mar 2020.

[46] Dina Abdelhafiz, Jinbo Bi, Reda Ammar, Clifford Yang, and Sheida Nabavi. Convolutional neural network for automated mass segmentation in mammography. *BMC Bioinformatics*, 21(S1), Dec 2020.

[47] Nils Bjorck, Carla P Gomes, Bart Selman, and Kilian Q Weinberger. Understanding batch normalization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman,

N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[48] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[49] Haoran Lu, Yifei She, Jun Tie, and Shengzhou Xu. Half-unet: A simplified u-net architecture for medical image segmentation. *Frontiers in Neuroinformatics*, 16:911679, 2022.

[50] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. Ghostnet: More features from cheap operations, 2020.

[51] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. In Sebastien Ourselin, Leo Joskowicz, Mert R. Sabuncu, Gozde Unal, and William Wells, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, pages 424–432, Cham, 2016. Springer International Publishing.

[52] Lukas Hirsch, Yu Huang, Shaojun Luo, Carolina Rossi Saccarelli, Roberto Lo Gullo, Isaac Daimiel Naranjo, Almir G. V. Bitencourt, Natsuko Onishi, Eun Sook Ko, Doris Leithner, Daly Avendano, Sarah Eskreis-Winkler, Mary Hughes, Danny F. Martinez, Katja Pinker, Krishna Juluru, Amin E. El-Rowmeim, Pierre Elnajjar, Elizabeth A. Morris, Hernan A. Makse, Lucas C. Parra, and Elizabeth J. Sutton. Radiologist-level performance by using deep learning for segmentation of breast cancers on mri scans. *Radiology: Artificial Intelligence*, 4(1):e200231, 2022.

[53] Dominik Müller, Iñaki Soto-Rey, and Frank Kramer. Robust chest ct image segmentation of covid-19 lung infection based on limited data. *Informatics in Medicine Unlocked*, 25:100681, Jul 2021.

[54] Mugahed A. Al-antari, Mohammed A. Al-masni, Mun-Taek Choi, Seung-Moo Han, and Tae-Seong Kim. A fully integrated computer-aided diagnosis system for digital x-ray mammograms via deep learning detection, segmentation, and classification. *International Journal of Medical Informatics*, 117:44–54, 2018.

[55] Khaoula Belhaj Soulami, Naima Kaabouch, Mohamed Nabil Saidi, and Ahmed Tamtaoui. Breast cancer: One-stage automated detection, segmentation, and classification of digital mammograms using unet model based-semantic segmentation. *Biomedical Signal Processing and Control*, 66, 2021.

[56] Timothy Cogan, Maribeth Cogan, and Lakshman Tamil. Rams: Remote and automatic mammogram screening. *Computers in Biology and Medicine*, 107:18–29, 2019.

[57] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L. Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data*, 4(1), Dec 2017.

[58] Mario Mustra, Kresimir Delac, and Mislav Grgic. Overview of the dicom standard. In *2008 50th International Symposium ELMAR*, volume 1, pages 39–44, 2008.

[59] Mammographic Image Analysis. Mammographic image analysis databases. https://www.mammoimage.org/databases/, 2011. Accessed: 30-10-2023.

[60] Inês C. Moreira, Igor Amaral, Inês Domingues, António Cardoso, Maria João Cardoso, and Jaime S. Cardoso. Inbreast: Toward a full-field digital mammographic database. *Academic Radiology*, 19(2):236–248, 2012.

[61] Adam Jaamour, Craig Myles, Ashay Patel, Shuen-Jen Chen, Lewis McMillan, and David Harris-Birtill. A divide and conquer approach to maximise deep learning mammography classification accuracies. *PLOS ONE*, 18(5), May 2023.

[62] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[63] Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3(1), May 2016.

[64] Dorothy Cheng and Edmund Y. Lam. Transfer learning u-net deep learning for lung ultrasound segmentation, 2021.

[65] Linting Guan and Yan Wu. Reduce the difficulty of incremental learning with self-supervised learning. *IEEE Access*, PP:1–1, 09 2021.

[66] Ramin Ranjbarzadeh, Nazanin Sarshar, Saeid Ghoushchi, Mohammad Esfahani, Mahboub Parhizkar, Yaghoub Pourasad, Shokofeh Anari, and Malika Bendechache. Mrfe-cnn: multi-route feature extraction model for breast tumor segmentation in mammograms using a convolutional neural network. *Annals of Operations Research*, 328, 05 2022.

[67] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. Data augmentation can improve robustness. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29935–29948. Curran Associates, Inc., 2021.

[68] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas, 2018.

[69] Andres Anaya-Isaza, Leonel Mera-Jimenez, Johan Manuel Cabrera-Chavarro, Lorena Guachi-Guachi, Diego Peluffo-Ordonez, and Jorge Ivan Rios-Patino. Comparison of current deep convolutional neural networks for the segmentation of breast masses in mammograms. *IEEE Access*, 9:152206–152225, Nov 2021.

[70] Wessam M. Salama and Moustafa H. Aly. Deep learning in mammography images segmentation and classification: Automated cnn approach. *Alexandria Engineering Journal*, 60(5):4701–4709, 2021.

[71] Hossein Soleimani and Oleg V. Michailovich. On segmentation of pectoral muscle in digital mammograms by means of deep learning. *IEEE Access*, 8:204173–204182, 2020.

[72] Nasibeh Saffari, Hatem A. Rashwan, Mohamed Abdel-Nasser, Vivek Kumar Singh, Meritxell Arenas, Eleni Mangina, Blas Herrera, and Domenec Puig. Fully automated breast density segmentation and classification using deep learning. *Diagnostics*, 10(11):988, 2020.

[73] Phillip Chlap, Hang Min, Nym Vandenberg, Jason Dowling, Lois Holloway, and Annette Haworth. A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 65(5):545–563, Jun 2021.

[74] P.Y. Simard, D. Steinkraus, and J.C. Platt. Best practices for convolutional neural networks applied to visual document analysis. *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.*, page 958–963, 2003.

[75] Eduardo Castro, Jaime S. Cardoso, and Jose Costa Pereira. Elastic deformations for data augmentation in breast cancer mass detection. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 230–234, 2018.

[76] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[77] Salome Kazeminia, Christoph Baur, Arjan Kuijper, Bram van Ginneken, Nassir Navab, Shadi Albarqouni, and Anirban Mukhopadhyay. Gans for medical image analysis. *Artificial Intelligence in Medicine*, 109:101938, 2020.

[78] Imran Ul Haq, Haider Ali, Hong Yu Wang, Lei Cui, and Jun Feng. Bts-gan: Computer-aided segmentation system for breast tumor using mri and conditional adversarial networks. *Engineering Science and Technology, an International Journal*, 36:101154, 2022.

[79] Sushovan Chaudhury, Alla Naveen Krishna, Suneet Gupta, K. Sakthidasan Sankaran, Samiullah Khan, Kartik Sau, Abhishek Raghuvanshi, and F. Sammy. Effective image processing and segmentation-based machine learning techniques for

diagnosis of breast cancer. *Computational and Mathematical Methods in Medicine*, 2022:1–6, Apr 2022.

[80] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[81] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248. Springer, 2017.

[82] Seyed Sadegh Mohseni Salehi, Deniz Erdogmus, and Ali Gholipour. Tversky loss function for image segmentation using 3d fully convolutional deep networks, 2017.

[83] Wolfram. Introduction to machine learning. https://www.wolfram.com/language/introduction-machine-learning/deep-learning-methods/, 2024. Accessed: 23-02-2024.

[84] Stephen M. Pizer, E. Philip Amburn, John D. Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B. Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, 39(3):355–368, 1987.

[85] Shuyi Li, Min Dong, Guangming Du, and Xiaomin Mu. Attention dense-u-net for automatic breast mass segmentation in digital mammogram. *IEEE Access*, 7:59037–59047, 2019.

[86] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

[87] Pavel Iakubovskii. Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch, 2019.

9 Appendix

9.1 Tables of full results

9.1.1 Hyperparameter Tuning Full Results

Starting LR	Minimum LR	Image Size	Batch Size	Total Epochs	Final Val Dice
0.001	1e-07	256	16	37	0.572

0.001	0.0001	512	4	15	0.476
1e-05	1e-06	256	32	23	0.516
0.001	0.0001	256	16	37	0.572
0.0001	1e-07	256	8	23	0.546
0.0001	1e-05	256	4	33	0.627
0.001	1e-07	256	32	35	0.537
0.001	0.001	256	16	47	0.563
1e-05	1e-05	256	4	22	0.362
0.001	1e-07	256	4	50	0.603
0.001	1e-06	256	32	35	0.537
0.001	0.0001	256	32	35	0.537
0.001	0.0001	256	4	50	0.603
0.001	1e-07	512	8	36	0.482
0.0001	0.0001	256	16	26	0.580
0.001	1e-05	256	4	50	0.603
0.001	0.001	512	8	32	0.071
0.0001	1e-06	256	16	24	0.602
0.0001	1e-07	512	8	28	0.505
1e-05	1e-06	256	16	15	0.438
0.001	1e-06	256	16	37	0.572
0.0001	1e-05	512	4	36	0.416
0.0001	0.0001	256	8	30	0.545
1e-05	1e-05	256	16	15	0.449
1e-05	1e-07	256	8	15	0.474
0.0001	1e-05	256	8	23	0.546
0.0001	1e-06	512	8	28	0.505
1e-05	1e-07	512	4	33	0.455
0.0001	0.0001	512	8	23	0.499
0.0001	1e-05	256	16	24	0.602
1e-05	1e-05	256	8	16	0.459
0.001	1e-06	256	8	37	0.481
1e-05	1e-07	512	8	30	0.472
0.001	1e-05	512	8	36	0.482
0.001	0.0001	256	8	37	0.481
0.0001	1e-06	256	32	24	0.559
0.0001	1e-06	512	4	36	0.416
0.001	1e-06	256	4	50	0.603
0.001	0.001	256	4	27	0.143
0.001	1e-05	512	4	15	0.476
0.0001	0.0001	256	32	31	0.569
0.001	1e-05	256	8	37	0.481

1e-05	1e-05	512	8	32	0.486
0.0001	1e-07	256	4	33	0.627
0.001	0.001	512	4	20	0.143
0.0001	1e-06	256	8	23	0.546
0.001	1e-07	256	8	37	0.481
1e-05	1e-05	256	32	25	0.347
0.0001	1e-07	512	4	36	0.416
0.0001	0.0001	256	4	16	0.550
0.001	1e-07	512	4	15	0.476
0.001	1e-06	512	8	36	0.482
1e-05	1e-06	512	8	30	0.472
0.001	0.001	256	8	50	0.512
0.0001	1e-07	256	32	24	0.559
1e-05	1e-05	512	4	33	0.452
1e-05	1e-07	256	32	23	0.516
1e-05	1e-07	256	16	15	0.438
0.001	1e-05	256	32	35	0.537
0.001	0.001	256	32	52	0.600
0.0001	1e-06	256	4	33	0.627
0.0001	0.0001	512	4	46	0.513
0.0001	1e-05	256	32	24	0.559
0.001	0.0001	512	8	36	0.482
1e-05	1e-06	256	8	15	0.474
0.001	1e-06	512	4	15	0.476
1e-05	1e-07	256	4	18	0.440
1e-05	1e-06	256	4	18	0.440
0.0001	1e-07	256	16	24	0.602
0.0001	1e-05	512	8	28	0.505
0.001	1e-05	256	16	37	0.572
1e-05	1e-06	512	4	33	0.455

Table 21: The full results of the hyperparameter tuning investigation.

9.1.2 Combining Data Augmentations Full Results

Hflip	Shear X	Shear Y	Translate	Rotate	Dice Score
✓	✓	✓	✗	✓	0.566
✗	✓	✓	✗	✓	0.531
✓	✗	✓	✗	✓	0.569
✓	✗	✓	✗	✗	0.574
✓	✓	✗	✓	✗	0.596

✗	✓	✗	✓	✓	0.602
✓	✗	✗	✓	✗	0.624
✓	✓	✗	✗	✓	0.595
✗	✗	✗	✗	✗	0.626
✗	✗	✗	✗	✗	0.632
✓	✗	✓	✓	✗	0.618
✗	✓	✓	✓	✓	0.598
✗	✗	✗	✓	✓	0.582
✗	✓	✓	✓	✗	0.526
✗	✓	✗	✓	✗	0.586
✓	✓	✗	✓	✓	0.577
✗	✗	✓	✗	✓	0.629
✗	✗	✓	✓	✓	0.639
✗	✓	✓	✗	✗	0.570
✓	✓	✓	✓	✓	0.624
✓	✓	✓	✓	✗	0.588
✓	✗	✗	✗	✓	0.641
✗	✓	✗	✗	✓	0.586
✓	✗	✓	✓	✓	0.601
✗	✗	✗	✗	✗	0.694
✗	✗	✓	✓	✗	0.655
✓	✓	✓	✗	✗	0.648
✓	✗	✗	✓	✓	0.652
✓	✓	✗	✗	✗	0.617

Table 22: All the results from the combined data augmentation investigations.

9.2 Ethics Approval Document

University of St Andrews | FOUNDED 1413

School of Computer Science Ethics Committee

11 October 2023

Dear Sean,

Thank you for submitting your ethical application which was considered by the School Ethics Committee.

The School of Computer Science Ethics Committee, acting on behalf of the University Teaching and Research Ethics Committee (UTREC), has approved this application:

Approval Code:	CS17268	**Approved on:**	11.10.23	**Approval Expiry:**	11.10.28
Project Title:	Using Deep learning for the detection and segmentation of breast cancer from mammograms				
Researcher(s):	Sean Alger				
Supervisor(s):	David Harris-Birtill				

The following supporting documents are also acknowledged and approved:

1. Application Form

Approval is awarded for 5 years, see the approval expiry data above.

If your project has not commenced within 2 years of approval, you must submit a new and updated ethical application to your School Ethics Committee.

If you are unable to complete your research by the approval expiry date you must request an extension to the approval period. You can write to your School Ethics Committee who may grant a discretionary extension of up to 6 months. For longer extensions, or for any other changes, you must submit an ethical amendment application.

You must report any serious adverse events, or significant changes not covered by this approval, related to this study immediately to the School Ethics Committee.

Approval is given on the following conditions:

- that you conduct your research in line with:
  - the details provided in your ethical application
  - the University's Principles of Good Research Conduct
  - the conditions of any funding associated with your work
- that you obtain all applicable additional documents (see the 'additional documents' webpage for guidance) before research commences.

You should retain this approval letter with your study paperwork.

Yours sincerely,

Wendy Boyter

SEC Administrator