Development and Optimisation of Deep Learning Pipelines for Cancer Detection in Mammograms

Author: 220034672

Supervised by David Cameron Christopher Harris-Birtill

MSc Artificial Intelligence Dissertation, University of St Andrews

August 2024

Abstract

Breast cancer is the number one cancer found in women and is the most screened cancer worldwide[1][2]. Unfortunately, studies find radiographers misdiagnose one-fifth of screenings on average[3]. Therefore, recent advancements in artificial intelligence could be utilised as a tool to help improve diagnostic accuracy, but is not standard practice in industry yet[4]. Research in the area claims to have produced deep-learning pipelines producing accuracies approaching 100% that should be moved to production despite critical discussions around overfitting and reproducability[5][6]. This study tests these claims for reproducibility, and vital parameters are identified and documented to narrow the search space for the optimal pipeline and produce a hybrid model with highly generalisable results. The results show that the ShuffleNetV2 and AlexNet hybrid outperform several hybrid convolutional neural network models and the founding vision transformer models. ShuffleNetV2 and AlexNet produced consistent results across validation and test sets, scoring 67.06% and 66.57% respectively, out-performing comparable studies[7][8] thereby building a foundation for future work for datasets with greater size and diversity.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, David Harriss-Birtill, for allowing me to embark on this deep learning journey. His expertise on the subject matter was always visionary, and his continual support and guidance have been invaluable throughout this process. I could not have asked for more from a supervisor. I am also sincerely grateful to my academic advisor, Edwin Brady, for his continued favour and support. My thanks also go to Stuart Norcross and Jose Marques, who assisted me in setting up and running multiple high-performance computing units. Their expertise and willingness to help were crucial to successfully completing my experiments.

Further, I would like to thank St Andrews and the wider academic community for the wonderful institution each individual curates and their personal contributions to the discussion and creation of knowledge. I thank the university for awarding me a scholarship and providing the continued support necessary for this research. The resources and opportunities available have been appreciated and significant to my academic and personal growth. I promise I gave it my all and will continue to do so.

On a personal note, I am deeply thankful to my friends for their unwavering love, patience, and belief in me. Their support has been my foundation throughout this journey. Finally, this dissertation is dedicated to and inspired by those I have witnessed suffering from health conditions. You can have ninety-nine problems until you lose good health - then you only have one problem.

Declaration

I, Brandon Linnett, declare that this dissertation, *"Development and Optimisation of Deep Learning Pipelines for Cancer Detection in Mammograms"*, and the work presented in it are my own and have been generated by me as the result of my own original research.

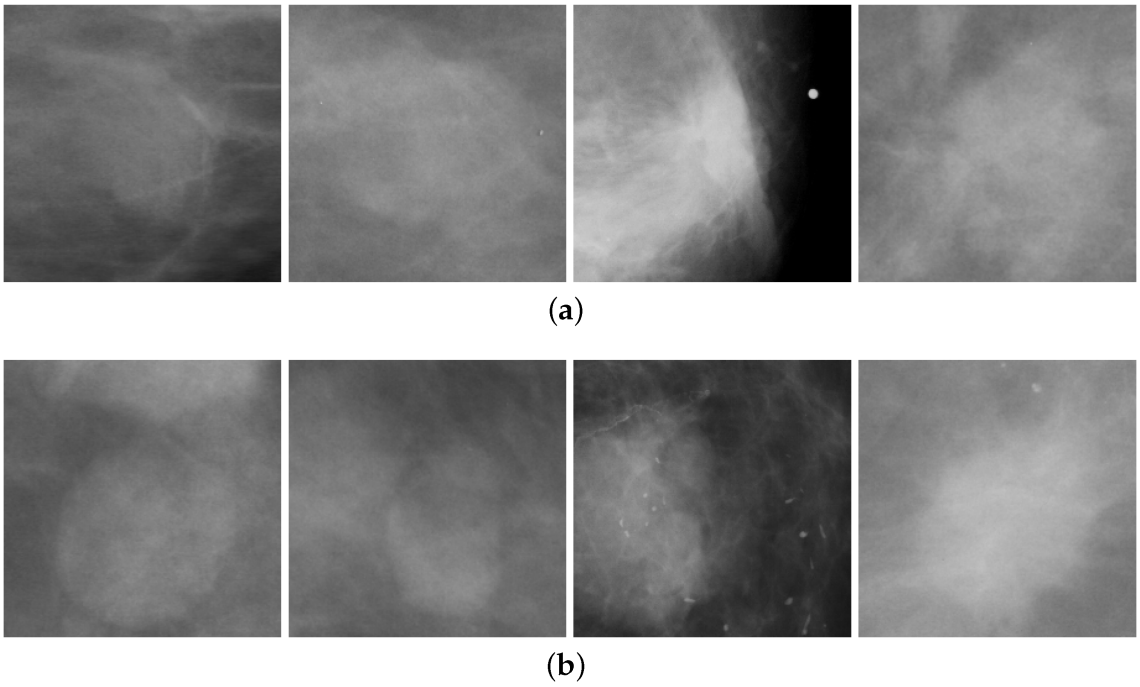
I confirm that:

- The word count is 14,436.
- This work was done wholly while in candidature for an Artificial Intelligence Masters degree at the University of St Andrews.
- Where any part of this dissertation has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- I have acknowledged all sources of help.
- Where the dissertation is based on work done jointly with others, I have made clear exactly what others did and what I contributed myself.
- This work has yet to be published before submission.

Contents

1 Introduction

Breast cancer is the 2nd most common cancer worldwide and the number one cancer in women, with around 2,296,840 recorded diagnoses and causing an estimated 670,000 deaths globally in 2022 [1]. Breast cancer can be hereditary, and therefore, individuals with associated gene mutations are more susceptible [9]. However, with this knowledge, mortality can be mitigated through enhanced screening, which increases the likelihood of survival [10] - hence, 39,000,000 scans are performed annually [2]. Sadly, even with scans, radiographers' accuracy in classifying scans is between 6191% [3]. For instance, Figure 1 shows four patients scans from the DDSM dataset, where two are benign and two are malignant - However, even for radiographers it may not be clear which are cancerous.



(a)

(b)

Figure 1: Cropped DDSM Mammograms - Two Benign and Two Malignant[11]

Radiographers are expensive and under time pressure to perform, and standard practices dictate that two radiographers confirm diagnoses of a mammogram[12]. Burdened healthcare systems may benefit from an assistant diagnostic tool utilising deep learning techniques by learning the underlying patterns in mammogram datasets. Deep learning has seen significant breakthroughs in the last decade, with convolutional neural networks correctly classifying 85% of 1.2 million images [13] and vision transformers scoring 88.5%[14].

Therefore, the overarching objective throughout this project is to classify mammograms accurately using deep learning. To achieve the objective, we develop intermediate objectives broken down into

primary, secondary, and tertiary objectives. The primary objectives are to conduct a comprehensive literature review to scope the subject and inform the methodology. A further primary objective is to develop a preliminary convolutional neural network and optimise fundamental components. The secondary objective is to develop and optimise advanced hybrid models with the best tested for generalisability on unseen data. Finally, the tertiary objective is to implement a vision transformer pipeline. The objectives here are chosen for their recent success in image classification tasks[14].

The structure of the report is as follows: First, acknowledging previous works seminal to the project, then the objectives achieved by the project, a deep dive into the literature which informs our project, the methodology as a result of the review, the presentation of our validation findings, then the presentation of the generalisability testing, followed by a discussion of the results and ultimately the conclusions of the project.

1.1 Previous Work

This paper presents research details regularly undocumented in much of the literature, though necessary. This research utilises the techniques from the highest-performing pipelines published to test reproducibility following machine learning best practices. Mainly, the aim is to reproduce methodologies prevalent in the literature, prioritising techniques from [5] and [6], who achieved 97.9% and 99% accuracy in breast cancer diagnosis. Notably, our work is aided by prior work by Adam Jaamour, Craig Macfadyen and Rhona McCracken, whose publicly available code aided some processing of the datasets[7] [15] [8].

2 Objectives

2.1 Description of the Problem

Many fields are at the precipice of realising a curated intelligence assistant [16], and paramount is the struggling healthcare service. Radiography needs more funding and staff[17]. Deep learning could assist in the crucial job of early detection and preventing radical surgeries.

The thesis proposes using deep learning, specifically convolutional neural networks (CNNs), to automate the detection of lesions in medical images of mammary carcinoma, as they are the most frequent architectures applied to medical imaging today [18]. The pre-processed data is passed through layers of configured models to fit the target variable and then backpropagated until a generalisable but accurate learning pipeline exists [19]. The high prevalence and mortality rates make Mammary Carcinoma a prime candidate for research.

2.2 Primary Objectives

- Conduct a comprehensive literature review of existing deep-learning methodologies, and their application to cancer detection in medical imaging.
- Pipeline Development for Cancer Detection: Design and implement a data pre-processing pipeline to handle the input datasets, including encoding, normalisation, augmentation, and splitting into training, validation, and test sets.

- Develop a preliminary convolutional neural network to classify mammograms, optimising key parameters: Loss functions, optimisers, and batch sizes.

2.3 Secondary Objectives

- Advanced Model Optimisation: Undertake a rigorous investigation to create and optimise an advanced deep learning pipeline. Specifically, to develop hybrid models aligning with prior implementation [5] and trial new hybrids.
- Perform generalisability tests

2.4 Tertiary Objectives

- Implement a vision transformer pipeline using the original vision transformers[14]

3 Literature Review

3.1 Convolutional Neural Networks

3.2 Introduction to CNNs

Convolutional Neural Networks (CNNs) have emerged as a pivotal tool in medical image analysis, offering significant advancements in the detection, classification, and segmentation of complex medical imagery [18]. The literature review guides a comprehensive examination of the current state of CNNs in medical image analysis, exploring their fundamental components, advantages, and challenges associated with their implementation.

The review is structured to guide the reader through the foundational concepts of CNNs, their specific benefits in medical imaging, and the limitations that researchers face. The review also delves into the various techniques employed to enhance CNN performance, including pre-processing strategies, data augmentation, and transfer learning. Furthermore, the review examines alternative architectures and methodologies exist, for a broader understanding of the topic. An overview of the datasets used in medical image analysis is present, highlighting their roles in advancing CNN research and application. Finally, the review identifies gaps in the literature and how they have informed the development of our research question: How can CNN architectures be utilised and optimised to improve the accuracy and efficiency of breast cancer detection?

3.3 Overview Of Convolutional Neural Networks in Medical Imaging

Convolutional Neural Networks (CNNs) are prominent in the literature for medical imaging classification tasks [18]. CNNs are promising as they map and learn the feature space presented [19], with adaptations approaching 100% accuracy in nodule detection for breast cancer [6][20][5][21]. CNNs offer significant advantages in automatic feature extraction and spatial relationship handling. However, they also face challenges, such as the need for large labelled datasets [13] and potential overfitting[22].

While CNNs have shown exceptional performance in mammography analysis, alternative approaches

such as transformers [23][24] and General Adversarial Networks [25] have also demonstrated promising results. These methods offer different strengths, such as improved handling of long-range dependencies in the case of transformers [26] or the ability to generate synthetic training data with GANs [27].

3.4 Historical Context of Convolutional Neural Networks

LeCun et al. [28] first introduced the CNN architecture in 1989 with the development of LeNet-5, an architecture primarily designed for handwriting recognition. Their early work laid the foundational principles of convolutional operations, weight sharing, and hierarchical feature extraction that would later become central to modern CNNs. Decades later, AlexNet's [13] deep architecture, combined with rectified linear units (ReLUs) [29] and dropout for regularisation [30], significantly outperformed traditional machine learning methods in a competition called ImageNet[31]. AlexNet achieved a top-5 error rate of 15.3% in the ImageNet Challenge, marking a paradigm shift in computer vision.

AlexNet's success spurred the application of CNNs to medical imaging, a domain involving the analysis of complex visual patterns. Pioneering studies, such as that by Cirensan et al[32], demonstrated the utility of deep CNNs in automating the detection of mitosis in breast cancer histology images, effectively surpassing the performance of conventional feature engineering approaches. Their work illustrated the potential of CNNs to revolutionise medical diagnostics by automating feature extraction, a process previously dependent on domain expertise and manual effort [33].

As computational resources and large annotated datasets became more accessible, for instance, due to advancements from Nvidia's GPU technology [34] and repositories such as The Cancer Imaging Archive [35], the deployment of CNNs in medical imaging proliferated. Notable architectures like VGGNet [36]and ResNet [37] further enhanced the depth and performance of CNNs, facilitating their adoption in more complex tasks, including multi-class disease classification, organ segmentation, and anomaly detection across various imaging modalities such as MRI, CT scans, and mammography [18].

CNNs' historical trajectory in medical imaging is marked by a series of pivotal innovations and applications, from early architectural developments to their current role in state-of-the-art classification systems. These advancements underscore CNNs' transformative impact on the field, enabling the automation of intricate image analysis tasks and setting the stage for further breakthroughs in medical artificial intelligence [38].

3.5 Overview of CNN Architecture and Components

CNNs are deep learning models designed for processing image data [19]. For the task, understanding the fundamental components of CNNs is critical for analysing the network's ability to detect subtle abnormalities in mammograms, such as tissue density variations, which are critical for early cancer detection and, therefore, in troubleshooting any technical problems for CNN performance.

At the core of CNN functionality are three critical architectural components: sparse connections, shared weights, and spatial hierarchies. Sparse connections ensure that only a portion of the input data is connected to each neuron, significantly reducing the parameters and computational load required to process high-dimensional images like mammograms [22]. Shared weights allow the same filter to be applied across different parts of the image, enhancing the model's ability to detect consistent features, such as edges and textures, across the entire image while also reducing redundancy [19]. Finally, spatial hierarchies enable CNNs to capture complex patterns by analysing the image at multiple scales, which is essential for understanding the intricate structures in medical images[39]. Figure 1 illustrates an example of this architecture.

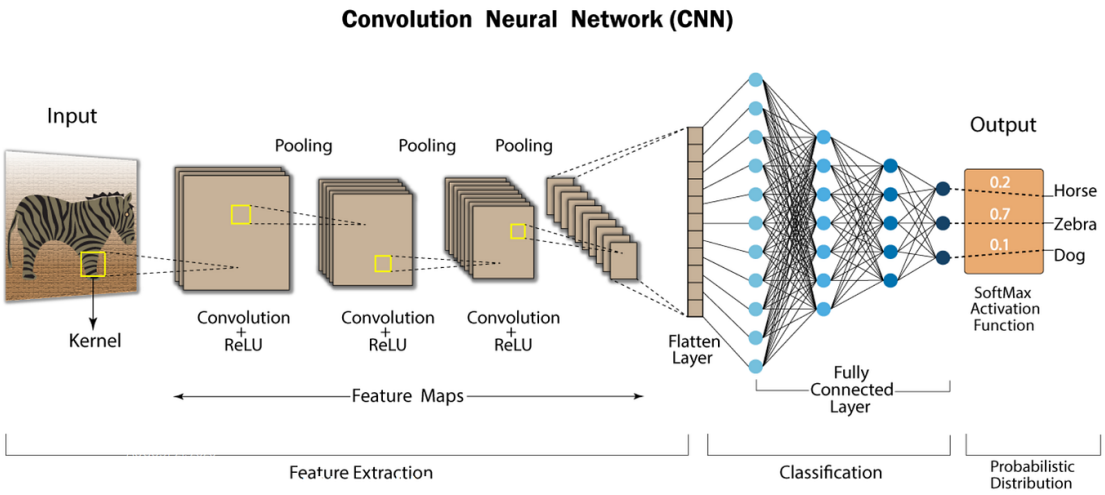


Figure 2: Convolutional Neural Network Architecture [40]

3.6 Convolutional Neural Net Components

Convolution Layers: Convolutional layers slide filters (small matrices) across the image, producing feature maps that capture various aspects of the image, such as edges, textures, and patterns[19][41]. Each convolution operation produces a new set of hidden variables called a feature map or channel. Additionally, these layers utilise shared weights, meaning that the filter can detect consistent features in the image regardless of geographical dispersion. Shared weights ensure abnormalities can be detected regardless of their spatial correlations [22].

Fully Connected Layers: After convolutional and pooling layers, fully connected layers perform the final classification by treating the features wholly. These layers take the high-level features extracted by the previous layers and output the class probabilities[13]. The dense layer will synthesise the features in the mammograms to classify the images as benign or malignant. Due to their dense connections, fully connected layers are prone to overfitting, requiring techniques like dropout and regularisation to improve generalisation [30].

In summary, CNNs are effective for medical image analysis [42] because they automatically learn relevant features from high-dimensional image data, efficiently handle spatial relationships, and perform robustly across classification tasks [13].

3.7 Benefits and Limitations of Convolutional Neural Networks in Medical Imaging

CNN's have become a cornerstone in medical imaging [18]. Their ability to adapt automatically and learn relevant features has resulted in many studies reporting high diagnostic accuracy - in many cases above 95%[42] [15][6]. However, despite the success and popularity of CNN's in medical imaging, accompanying drawbacks still require examination. This section examines the benefits, limitations, and any mitigation's.

Feature Learning and Noise: One significant advantage of CNNs is their ability to learn features directly from the images without manual feature extraction [22]. Feature learning is crucial in medical photographs where subtle differences signify crucial differentiation in patient conditions. Relevant features are highlighted by automatic adjustments to kernels, identifying components such as edges, densities, and patterns that are essential to the project. Automatic adjustments to kernels further ensure that the presence of noise in the image can still lead to reliable detection[43].

Handling High-Dimensional Data and Large Datasets:A typical 448x448 RGB image contains 602,112 input dimensions, representing a substantial challenge. However, CNN's gracefully handle high dimensionality as they process local image regions independently. Locality is a first-principles approach to breaking down a problem, reducing the parameter space and enabling CNN's to manage high-dimensional data at scale efficiently [38]. A demonstration of high-dimensionalality handling for classification is AlexNet, an architecture trained on 1,200,000 million images and tested on 150,000 images with an error rate of 15% [13]. A pertinent point, due to the usage of large datasets needed for this study.

Spatial Relationships: Capturing geographical relationships is critical to classification tasks as neighbouring pixels are often dependent. CNNs capture these spatial hierarchies by considering local pixel assemblies rather than processing individual pixels. Spatial awareness enables CNN's to capture patterns across the plane [22]. Spatial relationship capturing is essential in mammograms as lesions are matter groups, and CNN's would only be successful in detecting with this feature.

Automation of Routine Tasks: Finally, CNNs have the potential to automate routine diagnostic tasks, reducing the workload for radiologists and allowing them to focus on more complex cases. Automation can improve efficiency in clinical settings and reduce the likelihood of diagnostic errors due to human factors [44].

However, a fundamental problem in the general image classification endeavour is the large labelled datasets required for high performance [38][45]. Formalising the scale of this annotation problem, Macfadyen[15]composed a dataset of 4.5 million images from 102 individual datasets from the cancer imaging archive, enabling the team to achieve a test-set classification accuracy of 96%, exceeding prior similar investigations.

Arguably, annotation as a practicality problem concerning time and cost manifests dually as class imbalance. Class imbalance is a significant problem in cancer imaging archives due to the necessity of biopsies to confirm the diagnosis. Generally, benign (non-cancerous) samples are underrepresented[2]. Class imbalance occurs in mammography datasets as biopsies are not performed unnecessarily, which harms CNN classification performance if unaccounted for [46].

Limited data leads to another significant problem: overfitting, which limits the model's generalisability and clinical application [47]. Model overfitting occurs when the model can memorise the features of a training set and, therefore, scores highly in the performance metrics of the training data but does not perform well when testing for generalisability to unseen test data. As each body is unique, features in the test set may not have been present and, therefore, not factored into the model's weights during training.

The literature utilises numerous strategies to mitigate the consequences of limited data and the subsequent problem of overfitting. Regularisation techniques such as early stopping, weight decay, and dropout can prevent the model from learning noise that does not generalise well to unseen data [30]. Additionally, cross-validation, where the model is trained and validated against different subsets of the data, helps ensure the model performance is consistent across all subsets and is therefore helpful in identifying and reducing overfitting [48].

Another technique broadly applied in the medical imaging literature is data augmentation, which reduces overfitting by helping to learn more generalised features [49] and is shown to benefit performance metrics across all organs and modalities - though it is particularly effective for the heart, lungs and breasts[50]. Effective data augmentation techniques include geometrical transformations, erasing transformations, elastic transformations, pixel-level transformations, feature mixing, and General Adversarial Networks to increase dataset size and diversity artificially.

Buda[46] identified that the dominant strategy to address the class imbalance problem is oversampling the minority class. Duplicating the minority class until it levelled the majority class was found to be performance-aiding and did not cause overfitting, contrary to traditional machine learning techniques. Duplicating the minority class (the benign class) can address the class imbalance problem in the CMMD[51] and CBIS-DDSM[52] datasets, and the model can avoid biased predictions.

Transfer learning is another methodology to combat data insufficiency [53]. Pre-trained models such as AlexNet [13] can aid in learning features by transferring the model's weights to specific image classification tasks - effectively transferring feature recognition. Transfer learning in this manner becomes critical in establishing a model which can generalise well with limited data.

While CNNs are proving magical, that is also a downfall, as the chain of reasoning in decision-making is somewhat anonymised in the many layers and neurons [54], which is challenging to trust [55], especially in healthcare [56]. However, heat maps help identify the regions influencing the model's decision-making[57]. Furthermore, attention mechanisms can highlight areas of concern in the image, making the decision-making process more transparent and audit-able [26].

Despite challenges like the need for large labelled datasets, techniques like regularisation, data augmentation and transfer learning offer practical solutions, making CNNs the preferred choice for tasks like lesion detection in mammograms.

3.8 Alternatives and Enhancements to CNN Network Architectures

Following the discussion of the benefits and limitations of CNN architectures, this section examines alternative and enhancement methods that display exceptional results in our application domain. The section explores state-of-the-art applications in several sub-pockets of the broader machine-learning body of knowledge, thereby collecting a broader range of powerful tools. For a summation of the literature, table consolidations are append to this section.

Traditional Machine Learning and Optimisation Techniques
Traditional machine-learning (ML) techniques have shown significant utility in medical image analysis, particularly in smaller datasets [58][18]. Therefore, they still generate valuable insight for our task [59]. As a preface, performance metrics withheld in the dissections are generally exceptional, as with all included literature throughout the project. Exact performance metrics are reported in the Tables appended to the end of this subsection.

Nayak[23] developed a system for pathological brain detection that uses the Improved Jaya algorithm combined with an Extreme Learning Machine (IJaya-ELM) and orthogonal ripplet-II transform for feature extraction and classification. Highly sophisticated techniques include CLAHE for contrast enhancement [60], orthogonal discrete ripplet-II transform for polar coordinate conversion, Principal Component Analysis and Linear Discriminant Analysis for dimensionality reduction [61][62]. Similarly, Muduli [24] utilised the Fast Discrete Curvelet Transform (FDCT) [63] with Modified Particle Swarm Optimisation (MODPSO) [64] and Extreme Learning Machine (ELM) [65] for breast cancer detection. Khandezamin. [66] combined logistic regression for feature selection with the Group Method of Data Handling (GMDH) neural network [67][68], all scoring highly as recorded in Table 1. In summary traditional ML techniques, when highly optimised, offer success in the application domain. However, they require highly specialised feature engineering, losing generalisability[69].

Feature Selection and Bayesian Optimisation
Building on traditional machine learning techniques, Feature selection techniques combined with Bayesian Optimisation have been effective in improving breast cancer classification by reducing dimensionality and also improving computation time: Akkur[70] used feature selection methods like Relief [71], LASSO [72], and Sequential Feature Selection[73] with Bayesian Optimisation [74] for hyper parameter tuning. The Feature Selection and Bayesian Optimisation approach reduces the feature space and optimises their parameters to balance model complexity with predictive power to build a parsimonious model[70]. The combination of these techniques represented an evolution in the field, moving towards automated and efficient model development[75].

Naive Bayesian and Artificial Neural Networks
Naive Bayesian classifiers and artificial neural networks (ANN's) are helpful in breast cancer detection due to their simplicity and interpretability [76]. Naive Bayesian classifiers, introduced by

Duda and Hart[77], operate on the feature independence assumption, making them computationally efficient, particularly in scenarios with smaller datasets [78]. However, the feature independence assumption is a notable limitation, potentially leading to suboptimal model performance by not capturing feature relationships[79]. Karabatak [80] optimised a Naive Bayesian classifier using grid search for weight adjustment, demonstrating practical utility. Similarly, ANN's, which were first conceptualised by McCulloch and Pitts [81] and later developed further by Rosenblatt with the perceptron model [82], have been effective in breast cancer detection. Alshayeji[83] demonstrated the effectiveness of a shallow ANN with a single hidden layer for breast cancer detection. Shallow ANNs are valuable for their interpretability and lower computational demands than more complex deep learning models[84]. These studies highlight the continued relevance of these models, particularly in contexts where computational efficiency and model interpretability are critical. However, the simplicity of these models is a drawback, as they may not capture the complex patterns that more advanced models can[83][80].

Convolutional Neural Networks and Hybrid Architectures
As prior approaches limitations emerged, Convolutional Neural Networks excelled in mammography classification[18]. However, the challenges discussed in the benefits and limitations section have compelled researchers to explore hybrid architectures by combining various models and techniques to improve overall performance. Hybrid models utilise strength in diversity to optimise performance.

For example, Thangavel[6] combined ResNet and U-Net architectures for hierarchical feature learning and precise segmentation in digital mammograms[37][33] and integrated adaptive fuzzy median filtering for noise reduction[85] - demonstrating how high accuracy is possible in noisy medical images. Chakravarthy[86] used ResNet18[37] with a Crow-Search Algorithm[87] for optimisation in mammogram classification, employing deep feature extraction with its residual learning framework, mitigating the vanishing gradient problem in deep networks[37]. Arooj[88] also applied transfer learning using a customised AlexNet[13] model pre-trained on ImageNet[31] and fine-tuning on ultrasound and histopathology images with success.

Building on singular model transfer learning, Sahu[5] explored hybrid CNN models combining base models AlexNet[13], MobileNetV2[89], ResNet18[37], VGG16 [36] and ShuffleNet[90]. Developing several hybrid models: Hybrid 1 combined MobileNetV2 with ResNet18, leveraging MobileNetV2's efficiency and ResNet18's deep feature extraction capabilities. Hybrid 2 integrated VGG16 with ResNet18, combining VGG16's depth and simplicity with ResNet18's residual learning framework. Hybrid 3 paired ResNet18 with AlexNet, utilising AlexNet's pioneering deep learning architecture alongside ResNet18's modern features. Hybrid 4 combined ShuffleNet with AlexNet, where ShuffleNet's pointwise group convolutions and channel shuffling techniques[90] are paired with AlexNet to enhance computational efficiency. Finally, Hybrid 5 merged ShuffleNet with ResNet18, blending ShuffleNet's lightweight design with ResNet18's feature learning. These hybrid models demonstrated improved performance across mammogram and ultrasound datasets, underscoring the potential of such combinations to enhance accuracy and computational efficiency for mammogram analysis.

The studies outlined here and in Table 4 embody how convolutional neural networks can successfully use ImageNet's broad, generalisable classification weights and adapt to specific image classification tasks. However, dataset scale is still a significant problem for these architectures.

Table 1: Traditional Machine Learning and Optimisation Technique Summaries

Reference	**Pre-processing**	**Feature Extraction**	**Feature Reduction**	**Classification**	**Architecture**	**Evaluation**
Nayak et al. (2018)	CLAHE. 256 bins, clip limit $\beta = 0.01$.	O-DR2T. 2D DWT with Haar wavelet.	PCA and LDA. PCA+LDA: 2, PCA alone: 15.	IJaya-ELM.	IJaya-ELM: SLFN.	DS-66: 100%, DS-160: 100%, DS-255: 99.69%.
Muduli et al. (2021)	ROI Extraction: Cropping, resized to 128x128. Noise Removal and Enhancement: Standard methods.	FDCT-WRP. Decomposition into 4 levels.	PCA and LDA. PCA+LDA: 3, PCA alone: 28.	MODPSO-ELM.	MODPSO-ELM: SLFN.	MIAS: 100%, DDSM: 98.94%, INbreast: 98.76%.
Khandezamin et al. (2020)	Normalisation: Data normalised to [0, 1]. Feature Selection: Logistic regression.	FNA cytology characteristics.	Logistic Regression for Feature Selection.	GMDH Neural Network.	GMDH Neural Network.	WBCD: 99.4%, WDBC: 99.6%, WPBC: 96.9%.
Singh et al. (2023)	Data Cleaning: Handling missing values and duplicates. Feature Scaling: Normalisation and Standardisation.	Hybrid ESO-GSO.	Hybrid ESO-GSO.	KNN, LR, RF, DT, SVM, Ensemble Methods.	YOLO Detector, CNN, ResNet-50, InceptionResNet-V2.	Accuracy: 98.9578% (Random Forest with ESO).
Elkorany et al. (2022)	Removed incomplete samples. Normalisation: Min-max scaling or z-score normalisation.	30 features (WDBC), 9 features (WBCD).	WOA and DA for feature selection.	SVM with RBF kernel.	SVM with RBF kernel, WOA and DA Optimisation.	WBCD: 100% (50-50 split), 99.27% (10-fold CV). WDBC: 99.65% (50-50 split), 97.89% (10-fold CV).

Table 2: Naïve Bayesian and ANN Classifier Summaries

Reference	**Pre-processing**	**Feature Extraction**	**Feature Reduction**	**Classification**	**Architecture**	**Evaluation**
Karabatak et al. (2015)	Removed 16 records with missing values, Feature Scaling: Normalisation and Standardisation to have a mean of zero and a standard deviation of one	Features directly obtained from the WDBC dataset, including Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses	Weight optimisation using a grid search mechanism for the weighted Naïve Bayesian classifier	Naïve Bayesian Classifier: Standard NB Classifier, Weighted NB Classifier with optimised weights	Naïve Bayesian Classifier: Standard NB: Based on Bayes' theorem, assumes conditional independence between features, Weighted NB: Extends NB by incorporating weights for each attribute	Highest Performance Achieved: Sensitivity: 99.11%, Specificity: 98.25%, Accuracy: 98.54%
Alshayeji et al. (2022)	Removed records with missing values, Feature Scaling: Normalisation and standardisation, Data Splitting: 80% training, 20% testing, 5-fold cross-validation	WBCD: Visual assessment of FNA samples, values range from 1 to 10, WDBC: Extracted from digitized images of FNA samples using Xcyt software, 30 features per nucleus	No feature optimisation or selection algorithms used, all features tested in entirety	Model: Shallow ANN with one hidden layer, Input Layer: Number of inputs matched number of features, Hidden Layer: 100 neurons with ReLU, Output Layer: 1 neuron with sigmoid	Input Layer: Features from WBCD and WDBC, Hidden Layer: 100 neurons with ReLU, Output Layer: Sigmoid function for binary classification, Optimisation: Adam algorithm, Loss Function: Binary cross-entropy	WBCD: Sensitivity: 100%, Specificity: 99.72%, Accuracy: 99.85%, Precision: 99.69%, F1 Score: 99.84%, AUC: 99.86%, WDBC: Sensitivity: 99.59%, Specificity: 99.53%, Accuracy: 99.47%, Precision: 98.71%, F1 Score: 99.13%, AUC: 99.56%

Table 3: Feature Selection and Bayesian Optimization

Reference	**Pre-processing**	**Feature Extraction**	**Feature Reduction**	**Classification**	**Architecture**	**Evaluation**
Akkur et al. (2023)	Removal of ID numbers and class labels, Normalisation/Scaling: Z-score normalisation, Data Splitting: 10-fold cross-validation	WBCD: Extracted 30 features from images of cell nuclei (e.g., radius, texture, area, etc.), MBCD: Extracted 54 shape and texture features from ROIs (e.g., area, perimeter, GLCM features)	Relief (RF): Weighs features based on relationships, LASSO: Sets non-relevant feature coefficients to zero, SFS: Adds features sequentially until no improvement	Decision Tree (DT), Naive Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), Ensemble Learning (EL)	MATLAB 2020a with Statistics and Machine Learning Toolbox, Bayesian Optimisation (BO) for hyperparameter optimisation	WBCD: LASSO-BO-SVM achieved highest accuracy (98.95%), precision (97.17%), recall (100%), and F1-score (98.56%), MBCD: LASSO-BO-SVM achieved highest accuracy (97.95%), precision (98.28%), recall (98.28%), and F1-score (98.28%)
Dewangan et al. (2022)	Noise Removal: Wienmed filter (combination of Wiener and median filters), Image Cropping: Manual cropping to focus on relevant parts, Normalisation: Standardised resolution to 256x256 pixels, slice thickness <2 mm	ResNet18 architecture with pretrained weights from ImageNet, Input Image Conversion: Grayscale MRI images converted to RGB, Resizing: Resized to 224x224 pixels, Feature Output: Extracted 512 features using global pooling layer of ResNet18	Normalisation: Zero-center normalisation to scale values between 0 and 1	Model: BPBRW with Hybrid Krill Herd African Buffalo Optimisation (HKH-ABO), BP-BRW: Combines decision trees and recurrent neural networks, HKH-ABO: Hybrid optimisation combining Krill Herd Optimisation and African Buffalo Optimisation	Types of Layers: LSTM and dense layers for classification, Optimisation Algorithm: Adam optimiser, Loss Function: Categorical Cross-Entropy Loss, Regularisation Techniques: Dropout and early stopping	Accuracy: 99.6%, Precision: 99.9%, Recall: 99.9%, Sensitivity: 97.7%, Specificity: 92.5%, Error Rate: 0.12%

Table 4: Convolution Neural Networks and Hybrid Architectures

Reference	**Pre-processing**	**Feature Extraction**	**Feature Reduction**	**Classification**	**Architecture**	**Evaluation**
Thangavel et al. (2024)	Adaptive Fuzzy Median Filtering, Noise Removal, Normalization	ResNet and U-Net Architecture	-	Neural Network combining features from ResNet and U-Net	ResNet, U-Net	Accuracy: 99%, Precision: 98.6%, Recall: 99.01%, Specificity: 98.9%
Chakravarthy et al. (2022)	Noise Reduction, Image Cropping, Pectoral Muscle Removal	ResNet18 Architecture	-	Extreme Learning Machine (ELM) optimized with Crow-Search Algorithm	ResNet18	Accuracy: 97.193% (CBIS-DDSM), 98.137% (MIAS), 98.266% (INbreast), Kappa Coefficients: 0.956, 0.964, 0.974 respectively
Arooj et al. (2022)	Noise Removal, Image Cropping, Standardization	Customized AlexNet Model	-	Modified AlexNet with transfer learning	AlexNet	Accuracy: 99.4% (Dataset A), 96.7% (Dataset B), 99.1% (Dataset C), 100% (Dataset A2)
Sahu et al. (2023)	Data Extraction, Normalisation, Augmentation	Hybrid Models (various combinations of AlexNet, VGG16, ResNet18, MobileNetV2, ShuffleNet)	-	Hybrid CNN Frameworks	Various Hybrids	Accuracy: 99.17% (mini-DDSM), 96.52% (BUSI), 98.13% (BUS2), High precision, recall, specificity
Ragab et al. (2021)	Image Enhancement, ROI Extraction, Data Conversion	DCNN Architectures (AlexNet, GoogleNet, ResNet)	PCA	SVM Classifier with hybrid DCNN features	Hybrid DCNN-SVM	Accuracy: 97.9% (CBIS-DDSM), 97.4% (MIAS), AUC: 1.00 for both

3.9 Preprocessing Techniques

Pre-processing mammography datasets is critical in improving data quality and consistency, in turn enhancing model performance. Effective pre-processing techniques are essential for addressing variability and imperfections in imaging data (Huang et al., 1979; Pizer et al., 1987). This pre-processing section explores various pre-processing techniques, including data cleaning, integration, augmentation, feature extraction, feature reduction, and identifying regions of interest (ROI). This ensemble of techniques is critical for ensuring that CNNs and other models effectively learn and make predictions on the data.

Normalisation: Normalisation adjusts pixel values to a standard range, improving the consistency of input data across different images [28]. Techniques such as Min-Max scaling and zero-centre normalisation are commonly employed. For example, AlexNet utilised centring, where the mean pixel value is subtracted from each pixel [91][6]. Alternatively, Inception scaled pixel values to the range [-1, 1] [92]. Standard normalisation to the range [0, 1] is also widely used to ensure consistency across datasets [66].

Noise Reduction and Background Removal: Effective noise reduction techniques are essential for mammography images, which often contain significant noise and background artefacts. Techniques like adaptive median filtering [93], and wavelet denoising [94]help preserve critical image details while reducing noise. Background removal methods, such as cropping effectively eliminate extraneous background, enhancing image productivity[8].

Contrast Enhancement: Enhancing image contrast is particularly useful in mammogram scans, where the contrast between breast tissue and the background can be low. CLAHE (Contrast Limited Adaptive Histogram Equalisation) has been widely used to improve image quality without over-amplifying noise. CLAHE prevents over-enhancement by allowing user control over the clip limit value, making it a preferred choice for mammogram images [95].

Wiener Filtering: Wiener Filtering is a technique for noise reduction. Noise is reduced by considering local mean and variance, preserving details like edges while diminishing Gaussian noise [96]. This technique is suitable for mammograms as it diminishes noise without compromising critical structures [20].

Adaptive Median Filtering: As identified in the prior tables, adaptive median filtering is a popular choice in the literature for noise reduction and edge preservation. ADM improves image quality by dynamically adjusting the filter to local statistics and is, therefore, beneficial in mammography for preserving crucial structures without significant blurring [97]. This technique is often preferred in the literature for its balance between noise reduction and edge preservation [35].

Artefact Removal and ROI Extraction: Removing artefacts, such as radiographers' labels, and focusing on regions of interest (ROI) are crucial steps in preprocessing mammogram images. Artefact removal ensures that only relevant features are analysed, while ROI extraction enhances model accuracy and reduces training time by concentrating on the most critical parts of the images. Techniques like thresholding, which binarises images to isolate areas of interest, are effective for both tasks [98]. Morphological operations and k-means clustering [99] are commonly employed to isolate the breast area and identify lesions [100] [21]. Additionally, region-growing techniques are

often used to remove the pectoral muscle, leaving only the breast tissue for analysis [101]. These combined methods ensure that the model focuses only on the most relevant areas, improving overall performance.

Data Splitting and Balancing: Proper data splitting strategies, such as a 70% training, 20% validation, and 10% testing split, as wekk as five-fold cross-validation, ensure robust model training [102]. Data balancing techniques, including duplication and various augmentations like rotation, flipping, and scaling, help create a well-distributed dataset [21].

In conclusion, pre-processing techniques are vital in the performance of deep learning models, especially in mammography analysis. While these methods are helpful, research is needed to optimise image processing to ensure optimal noise reduction, particularly in clinical settings with high diversity of features and quality.

3.10 Data Augmentation

Data augmentation artificially increases dataset size, crucial for training deep-learning models with limited labelled data. Broad categories of augmentation techniques were addressed in the Limitations and Solutions subsection. Here, further details of some of the more straightforward techniques are provided.

Flipping and Rotations: Horizontal flipping is often more utilised than vertical flipping to preserve labels [103]. Losing the labels is a concern in some datasets, though labels will remain accurate in the DDSM and CBIS-DDSM datasets due to the accompanying csv's. Moreover, any rotation between 1 and 359 can augment the data, artificially create more data, and improve model performance [5][104].

Cropping: Cropping is a commonly employed augmentation technique [104][24] that effectively gives the model multiple perspectives of the same image. Randomly cropping sections of images (patches) with fixed aspect ratios is another effective augmentation technique. This method, used in models like GoogLeNet, ensures that the model learns from multiple perspectives of the same image.

Studies have demonstrated that these techniques significantly improve model accuracy. For instance, augmentations, including horizontal flips, zoomed sections, and ROI extraction, increased the U-Net model's accuracy from 70.26% to 85.96% [6]. In addition to the methods above, the literature explores other more advanced augmentation techniques. For instance, geometrical transformations, erasing transformations, elastic transformations, pixel-level transformations, feature mixing, and Generative Adversarial Networks (GANs) are employed to increase dataset size and diversity artificially [49]. These techniques benefit performance metrics across various organs and imaging modalities, with notable effectiveness in the heart, lungs, and breasts [50].

3.11 Datasets

CMMD

The Chinese Mammography Database (CMMD) provides a collection of mammographic images aimed at aiding in the diagnosis of breast cancer[51]. This dataset contains 5202 mammography scans from 1775 patients, all of whom have biopsy-confirmed benign or malignant breast tumours. The dataset includes each patient's Mediolateral Oblique (MLO) and Craniocaudal (CC) projections. CMMD helps address the limitations of previous databases by providing a large sample size, diversity in patient demographics, and detailed clinical data. Notably, the dataset has exclusion criteria: Patients with previous biopsies, breast prostheses and images with substantial motion artefacts. The CMMD images are in DICOM format, ensuring standardisation and ease of use in medical imaging applications. The dataset's comprehensive nature and inclusion of molecular subtypes make it a valuable resource for exploring predictive biomarkers and improving breast cancer diagnosis and prognosis (Wang et al. 2016; Cai et al. 2019). The additional molecular subtypes could be helpful in future research to see whether the additional labels improve inference.

CBIS-DDSM

The Curated Breast Imaging Subset of the Digital Database for Screening Mammography (CBIS-DDSM) is a curated dataset derived from the original DDSM dataset. CBIS-DDSM includes 10,239 images from 1566 patients selected and annotated by trained mammographers. The dataset is standardised and updated to DICOM format, with additional ROI segmentation and bounding boxes for more precise lesion identification[52]. CBIS-DDSM addresses the challenges researchers face using the original DDSM, such as non-standard compression formats and imprecise ROI annotations. CBIS-DDSM facilitates the development and evaluation of CADx and CADe algorithms by providing decompressed, high-quality images with accurate annotations. The dataset includes detailed metadata and is split into training and testing sets aiding the reproducibility of research findings[52].

3.12 Gaps in the Literature

Throughout our investigation into the literature several gaps have been identified. Number one, there is a tendency to omit exact parameters necessary for replication - particularly in the pre-processing of the data. The lack of transparency along the pipeline hinders validation of these results and disables future work building upon the work. Further, the review did not find results for specific augmentations impact upon the models validation accuracy, leaving a gap in understanding how these augmentations contribute. Finally, the research identified often reports near-perfect accuracy's in classification of mammograms, however there is little-to-no uptake in the marketplace, possibly due to concerns of over-fitting and the need for regularisation[5]. Therefore, this project aims to test these parameters and replicate claims produced.

4 Methodology

The methodology describes the initial approach to detecting mammogram lesions, the pipeline and model implementations, the parameter space explored and the evaluation metrics utilised to achieve the objectives.

4.1 Approach

The application of deep learning to the classification of mammograms produced promising results. Research reporting exceptional evaluation metrics are chosen to form the base of our investigation[6][5]. These authors detail their pipeline; however, the details could be more comprehensive. Therefore, the approach detailed herein attempts to recreate the successful pipelines and investigate the muted details. Cross-examination of the critical features from high-performance architectures was congregated in developing the pipeline as a preliminary success indicator. Previous reports by Jaamour and McCracken[7][8]identified a well-performing base model, which is also built upon. Specifically, the base models utilised for transfer learning are Inception V3, ResNet18, MobileV2, ShuffleNetV2, and AlexNet, which were all trained on ImageNet[92][37][89][105][13][31].

Before implementing the pipeline, developing a rudimentary version to produce timely investigations into different parameter alterations under time constraints inherent to a master's thesis was necessary. The work began with utilising the CMMD dataset. A critical point for the next chapter on validation results, as comparative work receives higher validation accuracy's on CBIS-DDSM compared to CMMD[8].

In summary, the objective of the research is primarily to investigate the claims produced and their respective pipelines' considerations: mainly, the models could be criticised for over-fitting and lacking regularisation techniques, as was identified in their respective considerations[6][5].

4.2 Pipeline

Data Preparation: Overview
This section presents an overview of the data preparation process. The pipeline details the downloading and processing of the data, data augmentation, and parameters explored. This project uses two datasets: CMMD and CBIS-DDSM. These files are initially downloaded in DICOM format, various filters are applied, and then the images are augmented. After, the data is split into training and testing sets. Finally the best model is tested on unseen data. The coming Chapter presents a detailed presentation of the pipeline, and an overview of the implementation is in Figure 3.

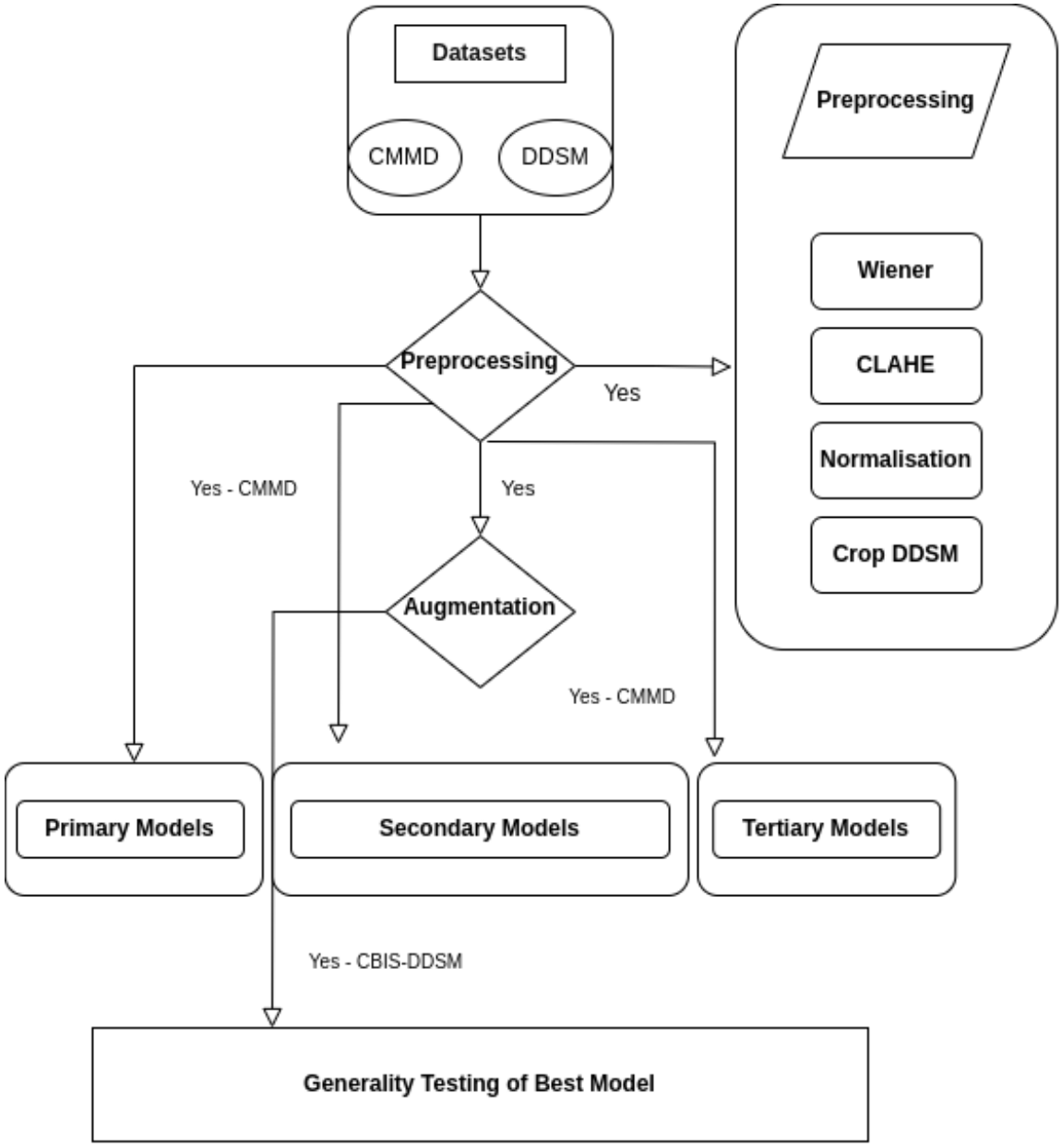


Figure 3: Overview of the Entire Pipeline

Data Loading
To import the datasets, the *get_ collection_ data.py* file is modified from prior work [15]. The script retrieves and saves metadata for our specified datasets from The Cancer Imaging Archive (TCIA). The metadata includes patient pathology information, image count, and imaging modality. Then,

the *tcia_ down load_ script* downloads the images by filtering the UIDs, checks if the series is downloaded already, and then downloads and extracts the zip file.

Data Processing
Next, the *data_ processing.py* file is employed. Running *data_ processing.py* applies the cleaning methods Wiener filtering, Contrast Limited Adaptive Histogram Equalisation, and normalisation to enhance the images for deep learning models. First, the script applies Wiener filtering to reduce noise while preserving details [20]. The *apply_ wiener* function on line 41 uses the *wiener filter* from *scipy* to operate. First, the function converts the images to floats, uses min-max normalisation, applies Gaussian smoothing, and applies the wiener filter. Wiener filtering is applied with various noise parameters [1.0, 0.1, 0.01], the affects of which are subject to investigation later, but, notably, artificially increases the dataset size.

After, the Contrast Limited Adaptive Histogram Equalisation (CLAHE) is applied with different clip limits [2.0, 3.0, 5.0] and tile grid sizes [8x8, 16x16] as a preferred choice for medical image processing [95] using *createCLAHE* function from the *cv2* library. CLAHE improves contrast in low-contrast areas to make subtle differences in tissue density more visible.

Finally, cropping application to the CBIS-DDSM dataset is applied due to substantial motion artefacts, which are detrimental to the detection process. Prior work by McCracken[8] was available online and applied for cropping the CBIS-DDSM images - viewable at *ddsm_ crop.py*.

Data Augmentation
After processing the images, the next stage of the pipeline is to augment the images to increase the dataset size artificially to improve generalisation, as was identified in the literature review as necessary [50]. Running *data_ augmentation.py* augments the images to create a further twenty images from the original, meaning the pre-processing and augmentation results in three-hundred-and-twenty generated images from a single image. The specific augmentations applied are:

- Horizontal and Vertical Flipping: Images are flipped left-right and up-down to simulate different orientations[103].
- Rotations: Images are rotated at angles of 45, 90, 135, 180, 225, and 270 degrees, with additional flips applied to the rotated images at 90 and 270 degrees[103].
- Scaling: Images are scaled by factors of 0.9 and 1.1 to introduce size variations[50].
- Translation: Images are translated by pixel shifts of (-10, 10) and (10, -10) along the x and y axes[50].
- Stretching: Images are stretched by factors of 1.2 and 0.8 along one dimension while maintaining the original aspect ratio[50].
- Shearing: Images are sheared by factors of 0.2 and -0.2, simulating changes in shape[50].

These augmentations are designed to reflect the literature and common variations in the real world, making the dataset more diverse and aiding model generalisation. The augmentations leverage tensorflow and efficient batch processing techniques to create a diverse and comprehensive dataset for

training deep learning models. The augmentations, including flipping, rotations, scaling, stretching, and shearing, hope to improve the eventual generalisation capabilities of the model, making it more effective in detecting lesions.

4.3 Model Architecture

The model architectures and associated hyperparameter trials explored to classify mammograms are outlined here according to our primary, secondary and tertiary objectives. Details of the model configurations, the hyperparameter optimisation tests, and the algorithms are presented in order to achieve the research objectives.

Primary Objective Model
The primary objective outlined in Chapter 1 was to produce a preliminary model and investigate various initial parameters. The manifest product of the objective is a baseline model with several scripts testing different loss functions, optimisation tools, and batch sizes—each initialising with predetermined sets of seeds. The primary objective manifestation serves as the foundation for subsequent experimentation by evaluating the capabilities of a standard convolution neural network architecture. The architecture used to achieve the primary objective is shown in Figure 3.

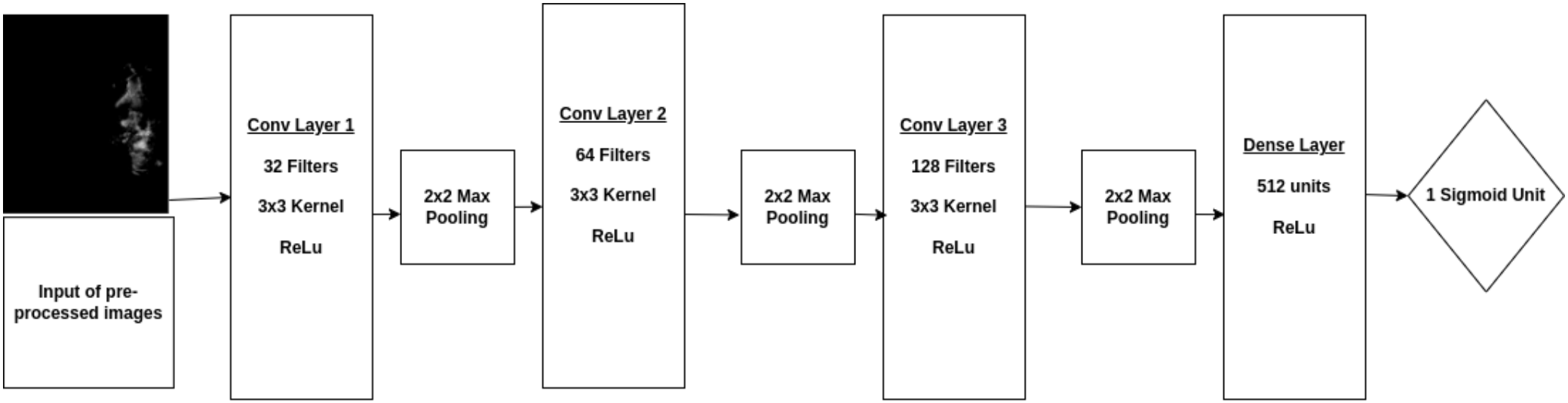


Figure 4: Overview of the Primary Objective Model

Model Architecture and Justification

- Input Layer The input size of (244,244) is useful for a balance between computation efficiency and performance, as demonstrated by AlexNet[13].
- Convolution Layers
  - Three Layers with Increasing Filters: Mirroring the approach in VGG and ResNet, where the number of filters increases with the depth, allows the model to capture evermore complex features [36][37]. Capturing complexity is essential in medical images as features with varying levels of abstraction exist.
  - ReLu Activation and Max Pooling: ReLu is widely utilised in the literature [13][36][37] due to mitigating the vanishing gradient problem in deep networks[29]. Ensuring gradients propogate through the architecture, ReLu ensures that training and convergence are efficient, providing high utility in medical images. Max pooling reduces the feature maps while retaining critical features[19]. Max pooling is helpful in medical imaging as the models can capture complexity in the training data that is unnecessary in achieving high diagnostic accuracy and avoids overfitting - improving model generalisability by focusing on the critical features.
- Dense Layer - 512 Dense Units with 50% Dropout: Including five-hundred-and-twelve dense units with ReLu activation accurately reflects established practices for medical imaging tasks where a balance between model complexity and performance is paramount. The choice of five-hundred-and-twenve units reflects the smallest number of units in the VGG16 architecture[36] to reduce the risk of overfitting on the subset of data used in validation. Further, the dense unit layer employs ReLu. ReLu introduces non-linearity, which enables the network to learn complex functions and is therefore chosen as the activation function in our fully connected layer. However, to reduce the risk of over-fitting, this architecture introduces a dropout rate of fifty per cent, meaning half of the units are inactive at every step. Dropout is essential to improve generalisability in limited datasets [30].
- Output Layer - Sigmoid Activation. Sigmoid manages the binary nature of our classification task by converting the dense layers output into a probability value and aligns with recommended deep learning practices outlined [19].

4.4 Primary Model Hyperparameter Testing

The above convolutional neural network was useful for primary parameter investigations. When developing deep learning models for mammogram classification, selecting appropriate loss functions, batch sizes, and optimisers is critical to improving performance. Each choice is grounded in theoretical principles and the literature's practical evidence.

Loss functions:

Loss functions are critical to model selection as they aim to minimise the error between predicted and actual values. Binary Cross-Entropy is a natural choice because it is specialised for modelling a binary outcome. Therefore, Binary cross-entropy is a valuable tool found in the literature for mammogram classification as the output hinges on a binary decision [6][86]. However, categorical

cross-entropy could aid in assessing scenarios where in-class variability is high, as it enables group differentiation of features, which could lead to higher classification accuracy[20].

Kullback-Leibler Divergence[106] was a unique application choice but one of interest. KLD can capture uncertainty in ambiguous cases by comparing predicted probability distributions with target distributions, which could help highlight cases requiring further examination. Finally, sparse categorical cross-entropy is tested to see if there is a speed increase while maintaining accuracy.

Optimiser Testing

Our hyperparameter testing extends to optimisers, initially using Stochastic Gradient Descent (SGD)[107], a foundational optimisation strategy that is simple yet effective. Additionally, RMSprop[91] is utilised in our parameter testing because it expands upon SGD by implementing adaptive learning rates that mitigate vanishing and exploding gradients. Finally, Adam[108], an optimiser in the lesion detection literature [83][6], builds upon RMSprop by integrating momentum, resulting in more better convergence across a broader search space. By systematically testing these optimisers, our study understands the effect of different optimisers on lesion detection.

Seed and Batch Sizes

The initial model explores the effects of different seeds and batch sizes on classification accuracy. Testing different seeds develops an understanding of the effect of different random initialisations of weights and other stochastic processes that affect the model's performance. Further, batch sizes are tested, determining the number of training images in each step which affects the updated gradients.

Augmentation and Pre-processing Testing

Next, the investigation sought to understand the specific impacts of various augmentations and pre-processing parameters on validation accuracy, as these artificial methods of increasing dataset size could be helpful for a project with the scope in computational resources, space and time. The series of augmentations tested are listed here:

- Original
- Original Flipped Horizontally
- Original Flipped Vertically
- Rotated 180 degrees
- Rotated 270 degrees
- Rotated 270 degrees Flipped Horizontally
- Rotated 90 degrees
- Rotated 90 degrees Flipped Horizontally

- Translated 300 x 0 y
- Translated Scaled 102 percent
- Translated Scaled 105 percent
- Translated Scaled 95 percent
- Translated Scaled 98 percent
- Translated Sheared -2 percent
- Translated Sheared 2 percent
- Translated Sheared -5 percent
- Translated Sheared 5 percent
- Translated Stretched 102 percent
- Translated Stretched 105 percent
- Translated Stretched 95 percent
- Translated Stretched 98 percent

The testing methodology involved combining each augmented version with the original dataset. This approach allowed for the evaluation of how each specific augmentation affected the accuracy of the validation. Tests were conducted across the same seeds and folds for consistency, enabling the isolation of each augmentation's impact on the model's performance. Maintaining consistent seeds and folds ensured that any changes in validation accuracy could be attributed directly to the augmentation technique rather than variations in data splitting or initialisation.

Finally, the investigation tested different pre-processing parameters for the CLAHE and Wiener filtering applications to the mammograms. Specifically, testing clip limits [2.0, 3.0, 5.0], noise level [0.01, 0.1, 1.0] and tile grid size [8x8, 16x16]. The following list provides the details of each variation tested, with CL referring to the clip limit, TG the tile grid size and WN the wiener noise level:

- **CL3.0_TG8x8_WN0.01**
- **CL3.0_TG8x8_WN1.0**
- **CL2.0_TG16x16_WN0.01**
- **CL5.0_TG16x16_WN0.01**
- **CL5.0_TG8x8_WN1.0**
- **CL5.0_TG16x16_WN0.1**
- **CL3.0_TG16x16_WN1.0**

- **CL2.0_TG8x8_WN0.01**
- **CL5.0_TG16x16_WN1.0**
- **CL2.0_TG16x16_WN0.1**
- **CL2.0_TG8x8_WN1.0**
- **CL5.0_TG8x8_WN0.01**
- **CL3.0_TG8x8_WN0.1**
- **CL2.0_TG8x8_WN0.1**
- **CL3.0_TG16x16_WN0.01**
- **CL2.0_TG16x16_WN1.0**
- **CL5.0_TG8x8_WN0.1**
- **CL3.0_TG16x16_WN0.1**

These variations were tested individually, allowing the model to find the best parameters for pre-processing mammogram data with Wiener filtering and CLAHE. The investigation aims to find the optimal way to reduce noise while preserving critical details necessary for high performance.

4.5 Secondary Objective Architectures

The secondary objective outlined in the introduction was to optimise an advanced deep-learning pipeline. The project utilises multiple base models and combines them to create a hybrid architecture which uses transfer learning to achieve secondary objective. Each hybrid model combines the strengths of multiple architectures, employing adaptability throughout the model to enhance the probability of an optimal detection outcome.

Generally, the approach tries to mimic that of Sahu[5] by using an initial extraction phase from model one to detect features. The generated feature map is passed to a primary classifier that evaluates the features captured. The second model is activated if the probability generated is beneath a classification threshold parameter. The second model also extracts features, refining the feature-capturing process. Next, the feature maps are combined with a conditional weight placed dynamically on the second model's feature set. The integrated set of features is passed to a series of fully connected layers, further learning the information in the input to produce a final classification prediction. The general pseudocode outlining this process is below; after, an example diagram for the ShuffleNet and AlexNet Hybrid architecture is present.

Algorithm 1 Hybrid CNN Algorithm for Breast Cancer Detection with Detailed Training and Evaluation - ShuffletNet & AlexNet Example

Input: Image x, Hyperparameters θ, Number of Cross-Validation Folds k
Output: Final Classification Prediction y, Evaluation Metrics
Phase 1: Dataset Preparation and Preprocessing
Load dataset from directory
Apply transformations: Resize, Normalise, etc.
Split dataset into k folds using GroupKFold to avoid data leakage
Phase 2: Hyperparameter Initialisation
Initialise hyperparameters θ (e.g., dropout rate, learning rate, batch size, etc.) using Optuna or Keras-Tuner
Phase 3: Training Loop for Each Fold
for each fold i in k **do**
    Phase 3.1: Model Initialisation
    Initialise ShuffleNet and AlexNet models
    Customise final layers to remove the output layer
    Phase 3.2: Initial Feature Extraction (Model 1 - ShuffleNet)
    $F_1 \leftarrow$ ShuffleNet Features(x)
    Phase 3.3: Primary Classification
    $p \leftarrow$ Sigmoid(Linear Layer(F_1))
    if $p <$ Conditional Activation Threshold(τ) **then**
        Phase 3.4: Conditional Feature Extraction (Model 2 - AlexNet)
        $F_2 \leftarrow$ AlexNet Features(x)
        Phase 3.5: Feature Integration with Dynamic Weighting
        $F_{\text{combined}} \leftarrow F_1 \cup (\omega \cdot F_2)$ ▷ Where ω is the conditional weight
    else
        $F_{\text{combined}} \leftarrow F_1$ ▷ Only use features from ShuffleNet
    end if
    Phase 3.6: Final Classification
    Pass F_{combined} through fully connected layers to obtain final prediction y
    Phase 3.7: Model Training
    Compute loss using Binary Cross-Entropy (BCE) Loss
    Update model parameters using an optimiser (Adam, RMSprop, or SGD) and learning rate scheduler
    Track training accuracy and loss
    Phase 3.8: Model Validation
    Evaluate model on validation set
    Track validation accuracy, loss, and other metrics (e.g., ROC curve, Precision-Recall curve)
end for
Phase 4: Model and Hyperparameter Optimisation
Optimise model and hyperparameters using cross-validation results
Select best performing model configuration
Phase 5: Final Model Evaluation
Load best model configuration
Evaluate on the complete dataset using GroupKFold
Calculate and save final evaluation metrics (confusion matrix, classification report, etc.)
Generate and save plots (learning curves, ROC curve, Precision-Recall curve)
Phase 6: Model Saving and Deployment
Save the best model and hyperparameters to the specified directory
Ensure reproducibility by saving the environment details
Return Final Classification Prediction y, Evaluation Metrics

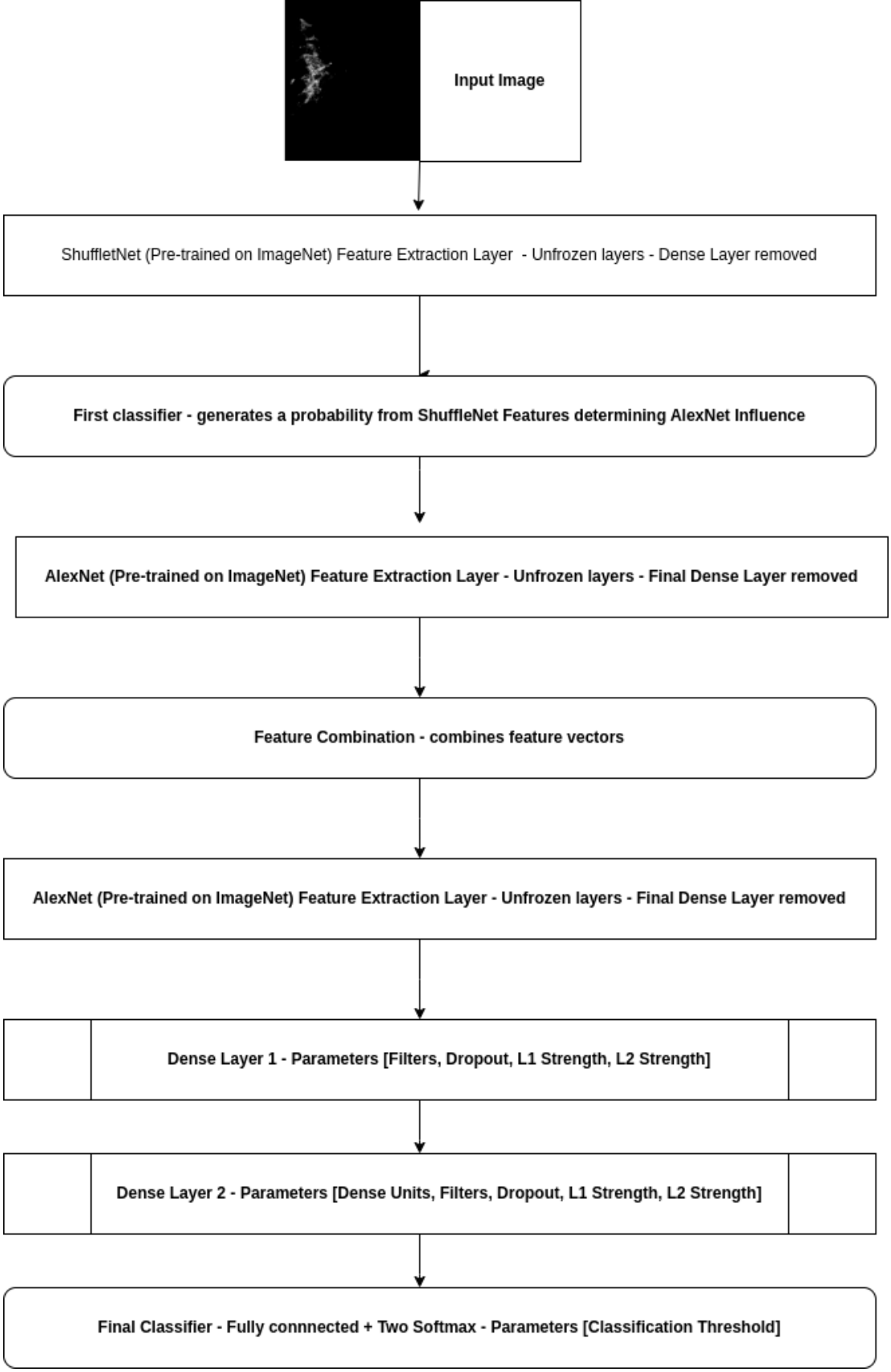


Figure 5: Secondary Objective Model Overview - Example of ShuffleNet & AlexNet

ShuffleNet & AlexNet Hybrid

The first hybrid model combines ShuffleNet and AlexNet, inspired by [5]. ShuffleNet uses two unique features, group convolution and channel shuffling, to reduce computation costs while maintaining accuracy [90]. On the other hand, AlexNet, the ImageNet competition winner, offers deep feature extraction. As a hybrid, the architecture balances computational efficiency and feature extraction capabilities as ShuffletNet is prime for mobile use, and AlexNet offers extensive feature extraction capabilities[13].

ShuffleNet & ResNet18

The second hybrid choice is also inspired by Sahu [5], who identified that the ShuffleNet & ResNet18 Architecture was second-best only to the hybrid above. Switching the secondary base model, the ShuffNet & ResNet18 hybrid switches AlexNet to ResNet18. While ShuffleNet maintains the computation efficiency, ResNet18 adds depth through eighteen layers with residual connections, mitigating the vanishing gradient problem and enabling subtle detail capturing [37].

InceptionV3 & ResNet18

Jaamour [7] and McCracken [8] inspire the third hybrid choice, as they found that the InceptionV3 base model performed well. InceptionV3 employs a broad range of filters with varying fields of vision to capture features at different magnifications [92], making InceptionV3 adaptable for capturing features in medical imaging. Combining with ResNet, which introduces residual connections for deeper networks, the study aims to identify whether adding base models to prior work aids classification. Particularly, the conditional feature weight on a deeper network may give further nuance to the network's decision-making process, but with subtle features given less weight.

ResNet50 & InceptionV3 & MobileV2

The selection of these base models is a novel choice and tests whether adding depth and complexity to the network aids in detecting intricate details and patterns in the data. Mainly, ResNet50 offers depth through fifty layers [37], InceptionV3 offers further detail recognition and space perception, while MobileNetV2 offers a lightweight addition for further depth with efficiency [89]. This selection of hybrid models is inspired by the prior work of Jaamour and Sahu et al. [7] [5] but notably is untested in the literature. Therefore this selection aims to answer whether adding a third model is beneficial for performance.

The hybrid models developed as part of our secondary objective aim to balance computational efficiency with deep feature extraction, making them highly effective for mammogram analysis. By combining different architectures and incorporating conditional features, these models aim to achieve high accuracy, aligning with the insights provided by recent studies and seminal works in the field.

4.6 Secondary Objective Hyperparameter Testing

As per the secondary objective to undertake advanced model optimisation, the research presents a rigorous investigation into the hyperparameters of the hybrid models. Throughout this subsection, hyperparameters are identified and briefly explained.

First, the investigation sought the best hyperparameters for each hybrid model. To do so, the investigation set up a hyperparameter space detailed in Table 5. For brevity, Table 5 examines the hyperparameter, the range of the hyperparameter explored, and the justification as to why it was essential to explore it.

Table 5: Hyperparameter Space and Justifications

Hyperparameter	**Range Explored**	**Justification**
seed	0 - 10000	Ensures reproducibility and tests robustness
dropout_rate	0.2 - 0.7[30] (step 0.1)	Prevents overfitting, improves generalisation
l2_strength	1e-6 - 1e-2 (log scale)	Controls model complexity, prevents overfitting
l1_strength	1e-6 - 1e-2 (log scale)	Induces sparsity, aids feature selection
learning_rate	1e-5 - 1e-2 (log scale)	Crucial for convergence and training speed
num_dense_units	128 - 1024[36] (step 128)	Affects model capacity and expressiveness
optimiser	adam, rmsprop, sgd[108][91][108]	Influences convergence speed and stability
batch_size	16 - 64 (step 16)	Affects training stability and generalisation
filters	32 - 128[36] (step 32)	Determines feature extraction capability
class_threshold	0.1 - 0.9 (step 0.1)	Balances precision and recall in classification
cond_act_threshold	0.5 - 0.9[5] (step 0.1)	Determines model two activation
cond_act_weight	0.1 - 0.9[5] (step 0.1)	Determines model two feature weight

4.7 Tertiary Objective - Vision Model Implementation

Transformers are an exciting recent development [26] and are updated for vision by Dosovitskiy in 2020[14]. This application utilises the original vision transformers[14] to achieve the tertiary objective of developing a vision transformer pipeline. Vision Transformer (ViT) implementations are utilised to exploit their ability to capture long-range dependencies within images. Capturing spatial relationships is especially beneficial for tasks such as mammogram analysis, where features such as the perimeter can be expansive. The specific ViT models used include ViT-B16, ViT-B32, ViT-L16, and ViT L32. These models differ primarily in the size of the patches they process and the depth of their architecture.

The Vision Transformer architecture, introduced by Dosovitskiy[14], represents a shift from traditional convolutional neural networks (CNNs) by employing a Transformer-based approach, initially developed for natural language processing. Critical components of this architecture include:

- **Image Patching**: Involves dividing images into fixed-size patches. Each patch is flattened and passed through a linear projection, reducing its dimensionality. Positional embeddings are a prerequisite for retaining the spatial structure of the image.
  - ViT-B16 and ViT-L16 use patch sizes of 16x16, capturing finer granularity.
  - ViT-B32 and ViT-L32 use patch sizes of 32x32, which reduce computational cost but provide a coarser representation.
- **Transformer Encoders**: Process the positional embeddings through multiple layers. They utilise multi-head self-attention mechanisms and feed-forward neural networks to capture relationships across the image.
  - ViT-B16 and ViT-B32 models consist of 12 layers of Transformer blocks.
  - ViT-L16 and ViT-L32 models consist of 24 layers, allowing for deeper processing and more complex feature representation.
- **Classification Head**: A classification token is prepended to the sequence of patch embeddings. This token aggregates information from all patches and is passed through a fully connected layer to produce the final classification output.

The vision transformer models were selected based on their demonstrated effectiveness in various image recognition tasks. Additionally, prior work demonstrates when comparing vision transformers to convolutional neural network models they perform on par in mammography[109]. Therefore, the investigation builds upon this and will compare the differences between vision transformers and hybrid model approaches.

4.8 Technology

Several technologies were critical throughout the project: Tensorflow, Pytorch, Vit_Keras, Optuna, and Keras-Tuner[110][111], which were used to implement and optimise the hybrid models. A significant proportion of the data handling was through the command line, as was necessary for working on the specified computers. Further, the source code was generated and deployed in a

Docker container utilising a Nvidia GeForce RTX 3060 Graphics Processing Unit provided by the university to speed up image processing during the preparation and validation phase. Further, for testing, the allocation of a DGX-1 machine with multiple Tesla V100s was granted for accelerated deep learning.

4.9 Reproducibility

Repeatability is a fundamental principle in scientific research. Reproducability ensures that findings can be independently evaluated by other researchers. In the context of building a pipeline for cancer detection, precisely documenting every factor that contributes to the model's development is critical. Documentation enables replication, comparison and development with transparency, therefore improving credibility and accelerates research. In Table 6 there is an exemplar itinerary conglomerated throughout the project that when documented substantiates the claims of a researcher and enables repeatability. Though notably the list is broad, and further work could produce a more precise matrix of each condition to enhance reproduction in this field.

Table 6: Comparison of Reproducibility Metrics across Projects

Reproducibility Point	Example Details to Document	This Project	Sahu et al. (2023)	Cantone et al. (2023)
Data Source and Access	Dataset sources and access details (e.g., public datasets, institutional databases) Data licences and usage permissions	✓	✓	✓
Patient IDs for Training, Validation, and Test Sets	Documentation of the specific patient IDs used in each data split to ensure exact reproducibility	✓	×	×
Data Preprocessing and Augmentation	Transformations applied (resize, normalisation, flipping, rotation, scaling) Preprocessing parameters used (e.g., CLAHE clip limits, Wiener filtering parameters)	✓	×	✓
Random Seed Documentation	Random seeds used for data splitting, weight initialisation, and augmentation processes	✓	×	×
Data Splitting Strategy	Detailed data splitting strategy (e.g., GroupKFold, seeds, patient-level splits)	✓	×	✓
Model Architecture and Configuration	Base models and configurations (ShuffleNet, ResNet, ViT models) Model layers, activation functions, dropout rates, hyperparameters	✓	✓	✓
Experimental Design	Justification for choices in model design, parameter settings, and experimental protocols	✓	✓	✓
Training Procedure	Hyperparameter tuning (ranges, tools like Optuna) Training regimen (epochs, batch sizes, learning rates, schedulers) Early stopping and checkpoint criteria	✓	×	✓
Weight Re-initialisation Across Folds and Splits	Procedure for re-initialising model weights for each fold and split in cross-validation	✓	×	×
Training Time and Resource Utilisation	Time taken for training (per epoch and total) Resource utilisation metrics (GPU/CPU usage, memory)	✓	×	✓
Evaluation Metrics	Performance metrics (accuracy, ROC-AUC, precision-recall) Cross-validation results (average scores, standard deviations)	✓	✓	✓
Final Model Selection Criteria	Criteria used to select the final model configuration (e.g., best cross-validation performance, lowest validation loss)	✓	✓	✓
Environment and Dependencies	Software and library versions (TensorFlow, PyTorch, scikit-learn) Hardware specifications (GPU model, memory capacity)	✓	×	✓
Model and Code Availability	Code repository (scripts, notebooks, clear instructions) Model weights (final trained model)	✓	×	×

4.10 Performance Metrics

The coming chapters on validation and testing explore the investigation's results. Several performance metrics are referenced: accuracy, precision, recall, and F1 Score. Additionally, each investigation is accompanied by graphs that formalise the concepts' numerical results visually. This subsection briefly details the equations producing the numbers and their interpretations to provide the reader insight into the meaning behind the presented results.

In the following equations:

- **TP** (True Positive) refers to the number of correct positive predictions.
- **TN** (True Negative) refers to the number of correct negative predictions.
- **FP** (False Positive) refers to the number of incorrect positive predictions.
- **FN** (False Negative) refers to the number of incorrect negative predictions.

Accuracy, the most straightforward metric, represents the proportion of correctly classified instances out of the total instances. Accuracy provides a broad overview of the model's performance.

$$\textbf{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{1}$$

Precision measures the proportion of true positive predictions among all positive predictions made by the model. Precision is essential when the cost of false positives is high, as it tells us how many predicted positives are correct. In mammography, precision scores indicate how many individuals have been incorrectly screened as having cancer.

$$\textbf{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{2}$$

Recall measures the proportion of actual positive cases that are correctly categorised. Recall is the paramount metric in cancer detection, as the cost of missing a true positive in prediction is high. A low recall score would indicate that many samples containing cancerous lesions are misdiagnosed as without pathology.

$$\textbf{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

The F1 score is the harmonic mean of precision and recall. F1 score provides a single metric that balances the two and is, therefore, useful as an overall effectiveness metric.

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

The validation and testing chapters provide confusion matrices and learning curves as a visual complement to the numerical metrics used in the investigations. Confusion matrices display a combinatorial grid of false negatives, true negatives, false positives, and true positives. Further, the presentation of learning curves is to illustrate how the model evolves during training, which aids in diagnosing model-specific problems such as overfitting. The performance metrics described here inform the reader's ability to evaluate the results presented in the following chapters.

5 Validation

5.1 Validation Introduction

Initially, the project hoped to explore a broad scope of parameters on the primary model to gain a broad and deep understanding of the components that impact validation accuracy. However, many were omitted due to time-constraints. For the secondary models, more rigorous investigations utilise Optuna and Keras-Tuner for optimisation over the larger hyperparameter space. The validation section first presents the investigations for the primary objective, followed by the investigations for the secondary and tertiary objectives.

5.2 Dataset Use for Validation

Throughout our main validation testing, a dataset of 5118 CMMD images using two augmentations of the original image: flipped horizontally and rotated 180 degrees representing a third of the population respectively. These images have undergone pre-processing by applying contrast limited adaptive histogram equalisation, wiener filtering and normalisation. The data is processed and split into 5 folds using scikit-learns GroupKFold function to ensure no data leakage between the training and validation set. Splitting the data into 5 folds for cross-validation is a common feature in the literature [6][86][20]. Additionally, the model is re-initialised between each fold to ensure no model weights cross over leading to over-confident performance metrics.

5.3 Batch and Seed Validation of Primary Model

An initial investigation was conducted to assess the effects of batch size and seed setting on model performance metrics. Various configurations were tested, including batch sizes [16, 32, 64, 128, 256] and seed values [0, 2000, 4000, 6000, 8000, 10000]. The results, summarised in Table 7, indicate that the highest validation accuracy of 65.59% was achieved with a batch size of 64 and a seed of 4000 in fold 5. Consistently higher validation accuracies were observed in fold 5, highlighting the importance of data selection during training.

Table 7: Best validation accuracy for each batch size with corresponding seed and F1-Score

Batch Size	Seed	Fold	Best Validation Accuracy (%)	F1-Score (Benign)	F1-Score (Malignant)
16	4000	5	64.75	0.35	0.60
32	6000	5	65.22	0.39	0.55
64	4000	5	65.69	0.40	0.56
128	8000	5	64.36	0.45	0.55
256	2000	5	62.60	0.49	0.52

The impact of different seed values and corresponding random initialisations and training paths was also examined, with the findings presented in Table 8. While the variance in validation accuracy across seeds was insignificant, seed 6000 produced the highest average validation accuracy of 63.94%, with a standard deviation of 1.515.

Table 8: Average validation accuracy and standard deviation for each seed value

Seed Value	Average Validation Accuracy Across Folds (%)	Validation Accuracy Std Dev
0	63.06	1.099
2000	62.75	1.007
4000	62.86	1.938
6000	63.94	1.515
8000	63.33	1.704
10000	62.95	1.609

Another observation from this investigation, as shown in Figure 4, with a batch size of 64, the training accuracy approaches 1 at epoch seven, but the validation accuracy plateaus at approximately 0.6. In contrast, Figure 7, which utilises a batch size of 256, shows the training curve reaching a plateau at epoch ten.

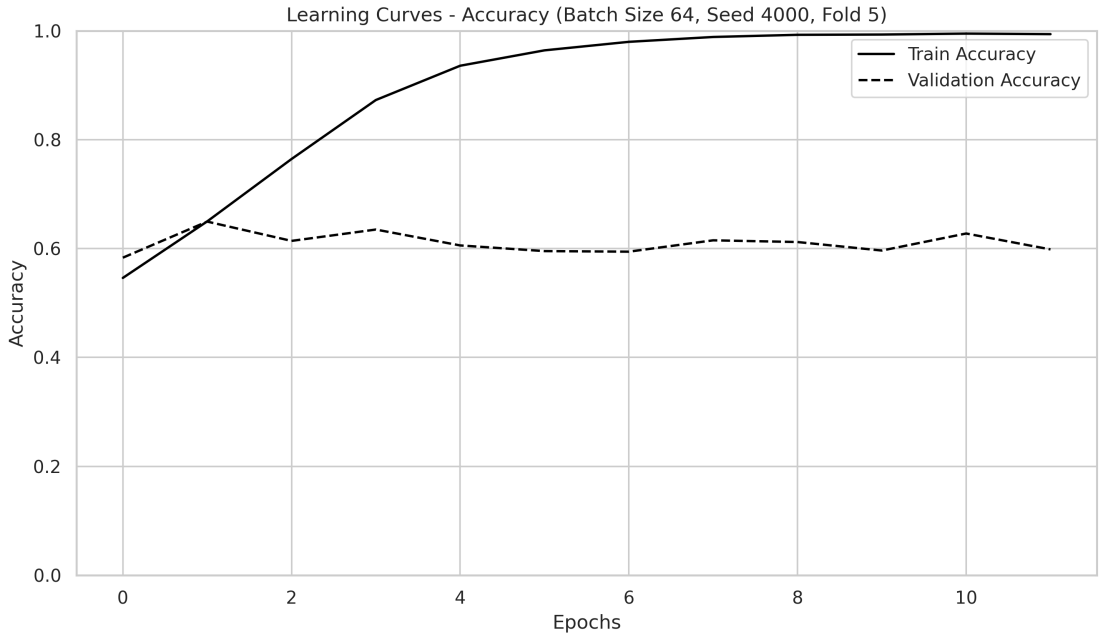


Figure 6: Learning Curve for Batch 64, Seed 4000

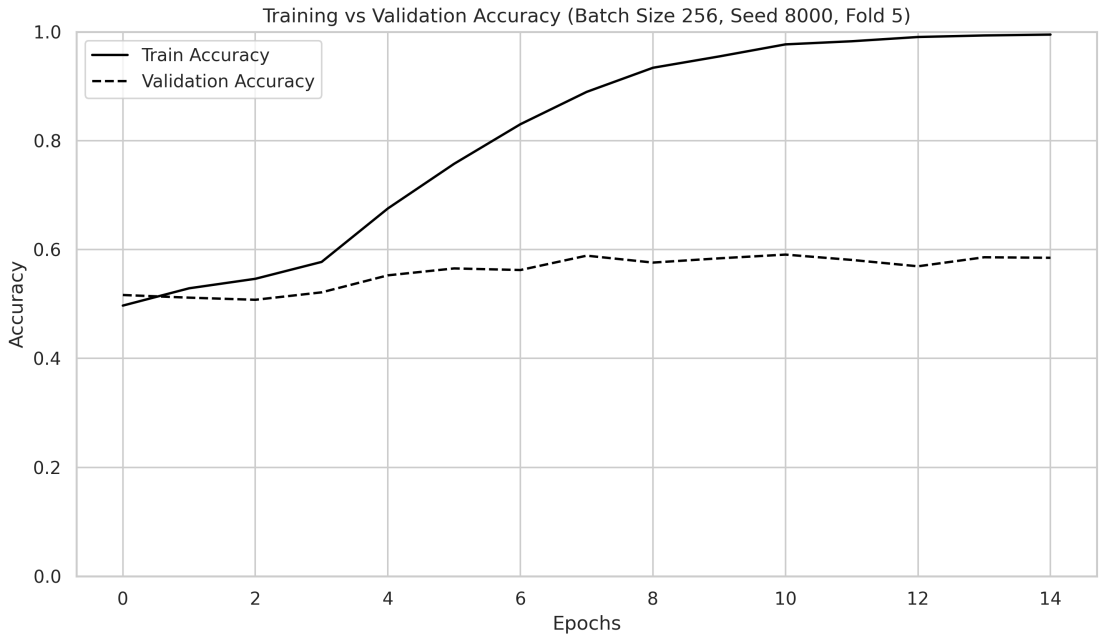


Figure 7: Learning Curve for Batch Size 256, Seed 8000

5.4 Loss Functions Testing of Primary Model

Subsequent investigation explored the performance of different loss functions commonly used in the literature. Contrary to expectations, the categorical cross-entropy loss function outperformed binary cross-entropy in our preliminary model, achieving a validation accuracy of 65.33%, as shown in Table 9. Other loss functions tested, such as Kullback-Leibler divergence (KLD) and Poisson, yielded lower validation accuracies.

Table 9: Best Validation Accuracy for Each Loss Function

Loss Function	Seed	Fold	Training Time (seconds)	Validation Loss	Validation Accuracy
binary_crossentropy	6000	5	2640.94	0.7040	64.84%
categorical_crossentropy	0	5	2873.14	0.6828	65.33%
kld	4000	5	4645.29	7.6349	52.63%
poisson	8000	5	4331.55	4.4455	55.31%

The confusion matrix depicted in Figure 8 summarises the model's performance on the validation set under the specified conditions. The results are as follows:

- **True Positives** (Malignant cases correctly classified as Malignant): 61.03%
- **True Negatives** (Benign cases correctly classified as Benign): 32.33%
- **False Positives** (Benign cases incorrectly classified as Malignant): 67.67%
- **False Negatives** (Malignant cases incorrectly classified as Benign): 38.97%

The dataset used for this evaluation contains 8% more malignant cases than benign cases, which is reflected in the distribution of predictions. The confusion matrix provides a clear breakdown of how well the model is distinguishing between benign and malignant cases under these conditions.

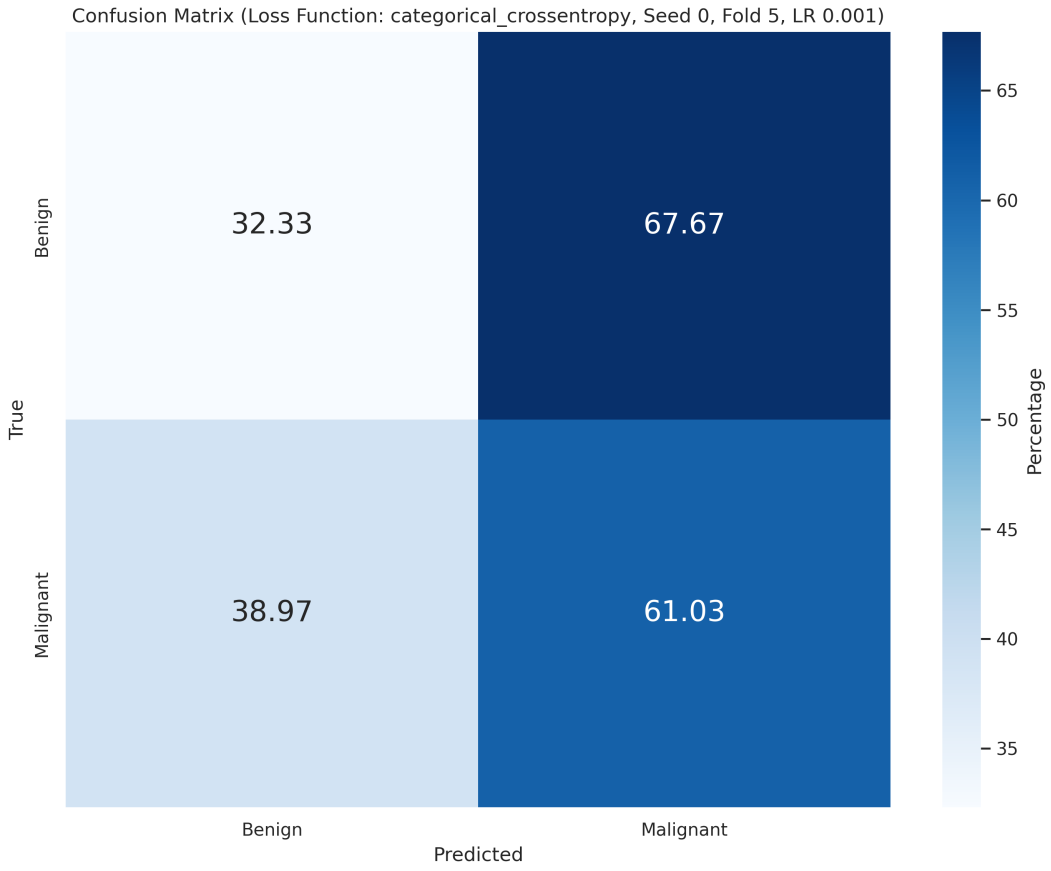


Figure 8: Confusion Matrix - Categorical Cross-Entropy, Seed 0, Fold 5 and Learning Rate 0.001

5.5 Optimiser and Learning Rate Validation of Initial Model

The final primary architecture investigation focused on evaluating the effectiveness of different optimisers and the impact of various learning rates on these optimisers. The results showed that the RMSprop optimiser consistently performed better across different seeds, folds, and learning rates, with the learning rate of 0.001 producing the best results for the dataset size of 5118 images. Detailed training times and validation accuracies for each optimiser are presented in Table 10.

Table 10: Best Validation Results for Each Optimiser

Optimizer	Learning Rate	Seed	Fold	Training Time (seconds)	Validation Accuracy (%)	Average Validation Accuracy Across Folds
Adam	0.001	2000	5	2670.46	65.92	62.10
RMSprop	0.001	10000	5	3590.20	67.20	64.56
SGD	0.001	10000	5	4137.11	66.70	64.11

5.6 Testing Pre-processing Parameters

The results in Table 11 indicate that the best performance is achieved with lower clip limits and smaller grid sizes, with CL2.0_TG8x8_WN1.0 and CL2.0_TG8x8_WN0.1 reaching the highest index of 1.0134. Although these configurations outperform others, there is little variation overall regardless of complexity or grid size.

Table 11: Configurations Sorted by Index (Highest to Lowest)

Configuration	**Index**
CL2.0_TG8x8_WN1.0	1.0134
CL2.0_TG8x8_WN0.1	1.0134
CL2.0_TG8x8_WN0.01	1.0094
CL3.0_TG8x8_WN1.0	1.0027
CL5.0_TG8x8_WN1.0	1.0027
CL3.0_TG8x8_WN0.1	1.0027
CL5.0_TG8x8_WN0.1	1.0027
CL3.0_TG16x16_WN0.01	1.0019
CL2.0_TG16x16_WN0.01	1.0019
CL5.0_TG16x16_WN0.01	1.0019
CL5.0_TG16x16_WN0.1	1.0019
CL3.0_TG16x16_WN1.0	1.0019
CL2.0_TG16x16_WN0.1	1.0019
CL2.0_TG16x16_WN1.0	1.0019
CL3.0_TG16x16_WN0.1	1.0019
CL5.0_TG16x16_WN1.0	1.0019
CL5.0_TG8x8_WN0.01	1.0002
CL3.0_TG8x8_WN0.01	1.0000

5.7 Testing Augmentations with the Primary Model

The results in Table 12 summarise the impact of adding various augmentations to the training data on the validation accuracy of the original dataset expressed as an index in comparison to using just the original data. The training script was designed to evaluate whether adding specific augmentations, as named in the variation, improves the validation accuracy compared to just the original dataset.

The investigation indicates that certain augmentations, particularly "Rotated 90 Degrees Flipped Horizontally" and "Rotated 90 Degrees," significantly improved the validation accuracy over the original dataset. However, other augmentations, such as "Original Flipped Horizontally," did not demonstrate an enhancement, yielding a lower validation accuracy. These findings suggest that specific augmentation strategies can be beneficial, while others may not contribute positively to model performance.

Variation	**Index**
Rotated 90 Degrees Flipped Horizontally	1.07
Rotated 90 Degrees	1.06
Original Flipped Vertically	1.06
Translated Stretched 98 Percent	1.06
Rotated 180 Degrees	1.05
Rotated 270 Degrees	1.05
Rotated 270 Degrees Flipped Horizontally	1.05
Translated 300 X 0 Y	1.05
Translated Sheared -2 Percent	1.04
Translated Sheared -5 Percent	1.04
Translated Stretched 102 Percent	1.04
Translated Scaled 105 Percent	1.03
Translated Sheared 5 Percent	1.04
Translated Scaled 98 Percent	1.03
Translated Sheared 2 Percent	1.03
Translated Stretched 95 Percent	1.02
Translated Scaled 102 Percent	1.02
Translated Scaled 95 Percent	1.02
Translated Stretched 105 Percent	1.01
Original	1.00
Original Flipped Horizontally	0.97

Table 12: Index Table Based on Validation Accuracy

5.8 Secondary Objective Validation Results

Before testing on unseen data, the project first uses a validation set to assess the models generalisability and find the best hyperparameters. For hybrid models combining ShuffleNetV2 with AlexNet, ShuffleNetV2 with ResNet18, and InceptionV3 with ResNet18, Optuna[111] was used for hyperparameter tuning. Optuna is an optimisation framework that efficiently explores the hyperparameter space using Tree-Structured Parzen Estimator[112]. Conducting multiple trials, each with a different set of hyperparameters, and prunes unpromising trials early to focus on the most promising configurations. Keras-Tuner[110] was utilised for the hybrid model combining ResNet50, MobileNetV2, and InceptionV3 by systematically searching for the best hyperparameters by running multiple training iterations with different hyperparameter using Random Search[113] with Hyperband[114].

The investigation identifies the best hybrid model as ShuffleNetV2 & AlexNet with a validation accuracy of 70.12% and F1-Score of 0.61 and 0.69 for Benign and Malignant, respectively. While InceptionV3 & ResNet18 scored 68.46% validation accuracy, the model performed poorly on the malignant class with an F1-Score of 0.61. ShuffleNetV2 & ResNet18 and ResNet50 & MobileNetV2 & InceptionV3 have slightly worse performance in classification. For a detailed breakdown of each hybrid model's classification report results, refer to Table 13. These results informed our selection for the generalisability testing, which will be presented in the following section.

Table 13: Classification Reports and Validation Accuracies

Model	**Validation Accuracy**	**Precision (Benign)**	**Recall (Benign)**	**F1-Score (Benign)**	**Precision (Malignant)**	**Recall (Malignant)**	**F1-Score (Malignant)**
ShuffleNetV2 & AlexNet	**70.12%**	**0.68**	**0.56**	**0.61**	**0.64**	**0.75**	**0.69**
InceptionV3 & ResNet18	68.46%	0.61	0.73	0.66	0.68	0.55	0.61
ShuffleNetV2 & ResNet18	67.97%	0.63	0.60	0.62	0.65	0.68	0.66
ResNet50 & MobileNetV2 & InceptionV3	65.98%	-	-	-	-	-	-

Table 14: Hyperparameters of the Models

Model	**Batch Size**	**Dropout Rate**	**Learning Rate**	**Dense Units**	**Optimiser**	**Filters**	**L1 Strength**	**L2 Strength**	**Classification Threshold**	**Conditional Activation Threshold**	**Conditional Activation Weight**
ShuffleNetV2 & AlexNet	**32**	**0.3**	2.21×10^{-5}	**768**	**adam**	**96**	1.03×10^{-5}	5.91×10^{-4}	**0.4**	**0.8**	**0.4**
InceptionV3 & ResNet18	32	0.6	1.35×10^{-4}	256	adam	96	3.55×10^{-4}	3.05×10^{-4}	0.9	0.5	0.9
ShuffleNetV2 & ResNet18	32	0.4	1.65×10^{-4}	512	adam	96	5.19×10^{-4}	2.10×10^{-3}	0.4	0.6	0.1
ResNet50 & MobileNetV2 & InceptionV3	48	0.3	8.13×10^{-4}	512	adam	96	2.95×10^{-5}	1.52×10^{-5}	0.4	0.6	0.7

5.9 Hybrid Models Hyperparameter Optimisation Results

The hyperparameter optimisation testing for various hybrid models, as presented in Table 12, revealed both similarities and differences across the architectures. Most models utilised a batch size of 32, with one exception using 48. All models consistently employed the Adam optimiser and 96 filters. However, dropout rates ranged from 0.3 to 0.6 across the models. Learning rates varied from 2.21×10^{-5} to 8.13×10^{-4}. The number of dense units differed among models, ranging from 256 to 768. L1 and L2 regularisation strengths showed variation across all models. The ShuffleNetV2 and AlexNet models had the lowest learning rate and highest number of dense units. These results found the best model to be tested for generalisability, though a deep discussion of the differences between architectures is reserved for the discussion chapter.

5.10 Tertiary Objective - Vision Transformer Results

The tertiary objective for this project was to develop a vision transformer pipeline. This subsection presents the validation performance for the transformers outlined in the methodology section. ViT-B16 performs the best among models that completed testing on time, with a validation accuracy of 67.29%, as shown in Table 15.

Table 15: Validation Accuracy Table

Model	**Seed**	**Fold**	**Validation Accuracy**
ViT-B16	2000	3	67.29%
ViT-B32	0	3	66.89%

The confusion matrices for ViT-B16 in Figure 9 reveal a bias towards predicting the malignant class, with an 83.09% accuracy in identifying malignant cases, compared to 16.03% of benign cases correctly identified. This imbalance is representative of all of the trials run, indicating that the vision transformers are heavily affected by class imbalance.

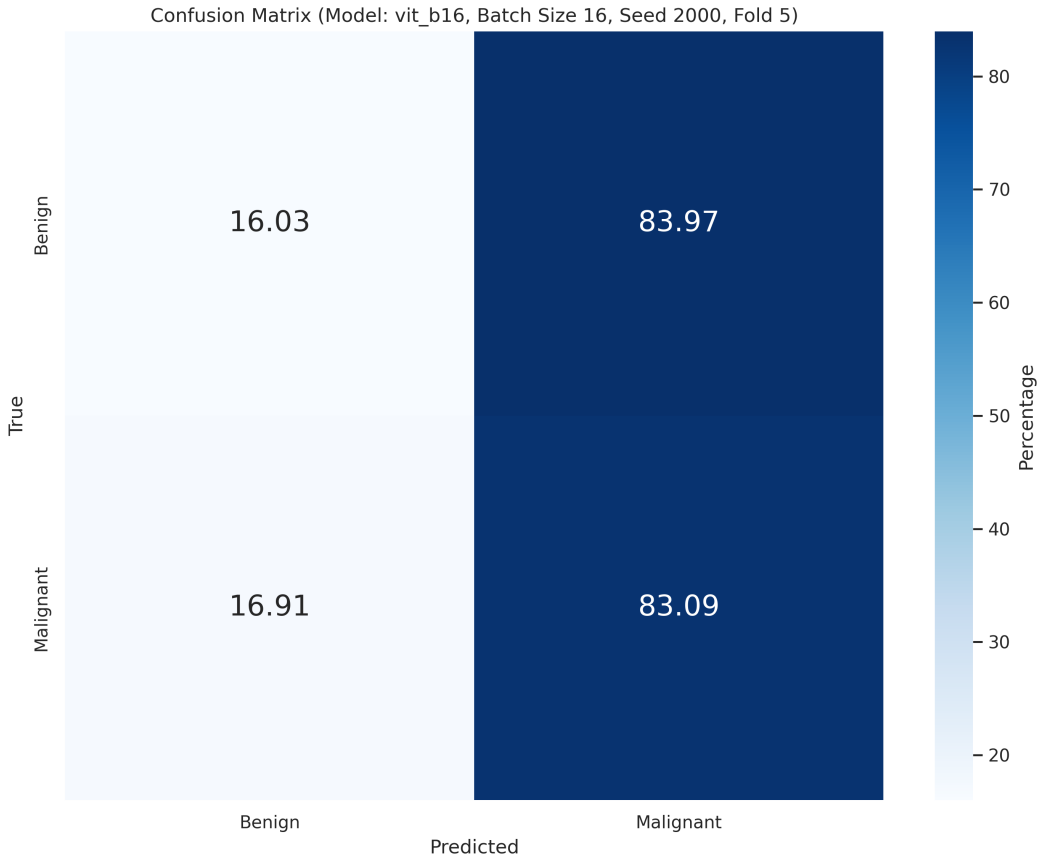


Figure 9: Confusion Matrix Example for ViT-B16

The learning curve for the ViT-B16 model in Figure 10 displays the training and validation accuracy, representing a pattern in the data this project generated. The general trend shows that:

1. Training Accuracy: The training accuracy starts around 60% and gradually improves to around 65-70% before plateauing. This slight improvement indicates that the model is learning, but plateus quickly, possibly due to the complexity of the task or insufficient training data.

2. Validation Accuracy: The validation accuracy curve peaks during the first epoch (here reported as 0) but then hovers below. The fluctuation in the validation accuracy suggests that the model does not learn generalisable patterns in the data well across training.

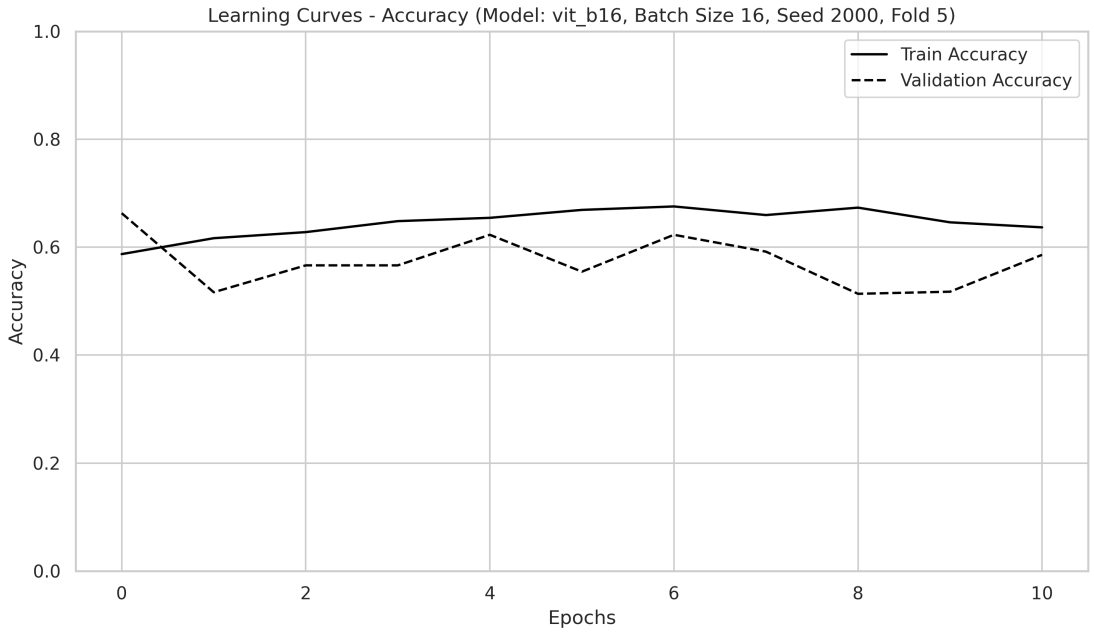


Figure 10: Learning Curve Example for ViT-B16

The learning curve represents a pattern across the transformers' learning curves the project tested, implying that the models consistently face challenges in learning effectively from the data. The performance outlined here could be due to the dataset's nature, model architecture, or training settings. Future work should experiment with different hyperparameters, architectures, or more data augmentation techniques to improve the model's generalisation ability. The discussion expands upon the optimisation of these models.

6 Testing for Generalisation

Throughout the project, the goal was to find a model that could be generalised across datasets. The validation investigations identified that the ShuffleNetV2 & AlexNet architecture was the most generalisable to unseen validation data. Based on that finding, the project presents further generalisation tests on a separate dataset - CBIS-DDSM - to check that the model produces consistent results across datasets. This generalisability testing is part of our secondary objectives and is designed to build trust in the metrics produced by the model.

Table 14 summarises the project efforts. The model achieves a test set accuracy of 66.57%, only 0.49% lower than the validation result on which the model parameters were tuned. The weighted F1-Score is precisely the same, demonstrating the consistency of the model's performance.

Model	**Accuracy (Validation Set)**	**Accuracy (Test Set)**	**Weighted Avg F1-Score (Validation Set)**	**Weighted Avg F1-Score (Test Set)**
ShuffleNetV2 & AlexNet	67.06%	**66.57%**	0.65	0.65

Table 16: Model Performance on Validation and Test Sets

Despite the consistency at the surface, there are significant differences in the makeup of these metrics. In the validation data, the model is significantly better at predicting true malignant cases, with a true positive rate of 75.55% but a 47.35% false positive rate as seen in Figure 11. In contrast, the test set confusion matrix (Figure 12) shows a dynamic change in the models behaviour from favouring malignant diagnosis to benign diagnosis. The true positive malignant rate drops to 46.12% and a high false negative rate of 53.88%.

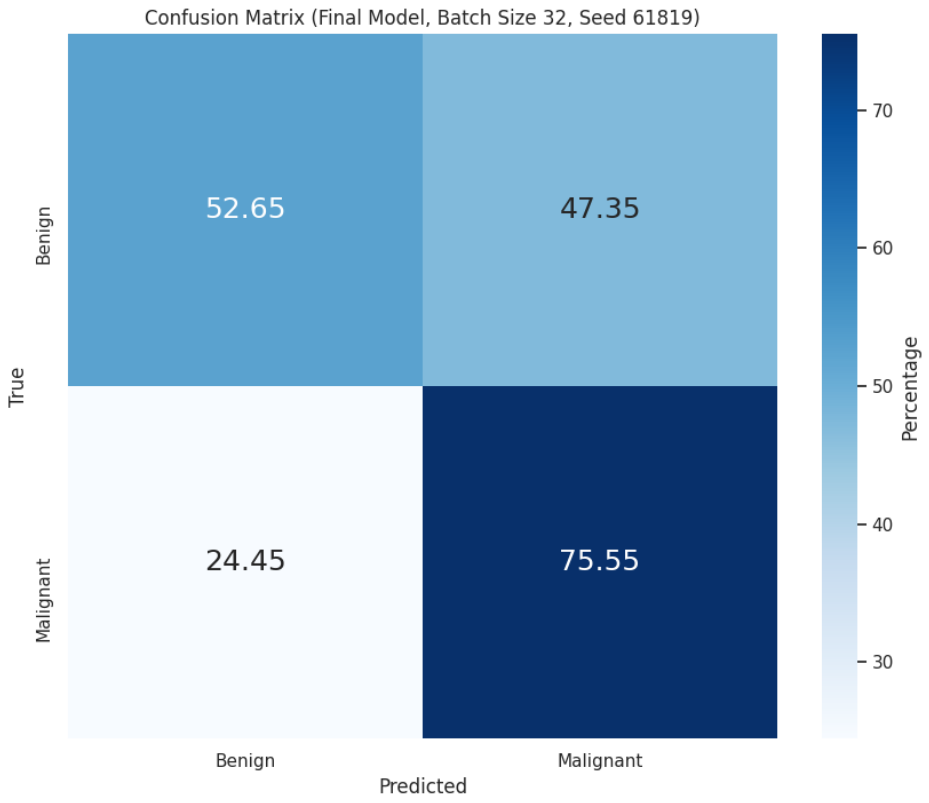


Figure 11: Confusion Matrix for CBIS-DDSM Validation Data

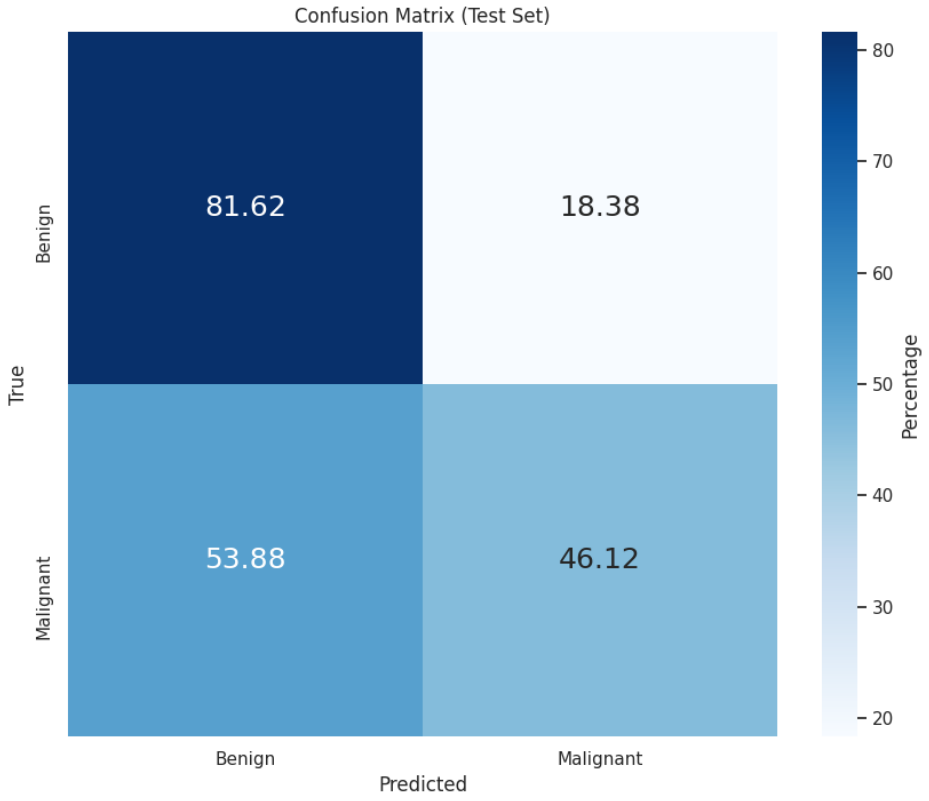


Figure 12: Confusion Matrix for CBIS-DDSM Test Data

The learning curve in Figure 13 represents the training and validation metrics across the 200 epochs where the model was training. The training phase employed two stages:

- **Epoch 1-100:** Only the dense layers of the model are trained, showing a gradual and close increase in training and validation accuracy, indicating the model is learning effectively without over-fitting.
- **Epoch 100:** ShuffleNetV2 layers are unfrozen, fine-tuning the model for feature extraction to our specific problem domain, providing further boosts to the validation accuracy but begins over-fitting.

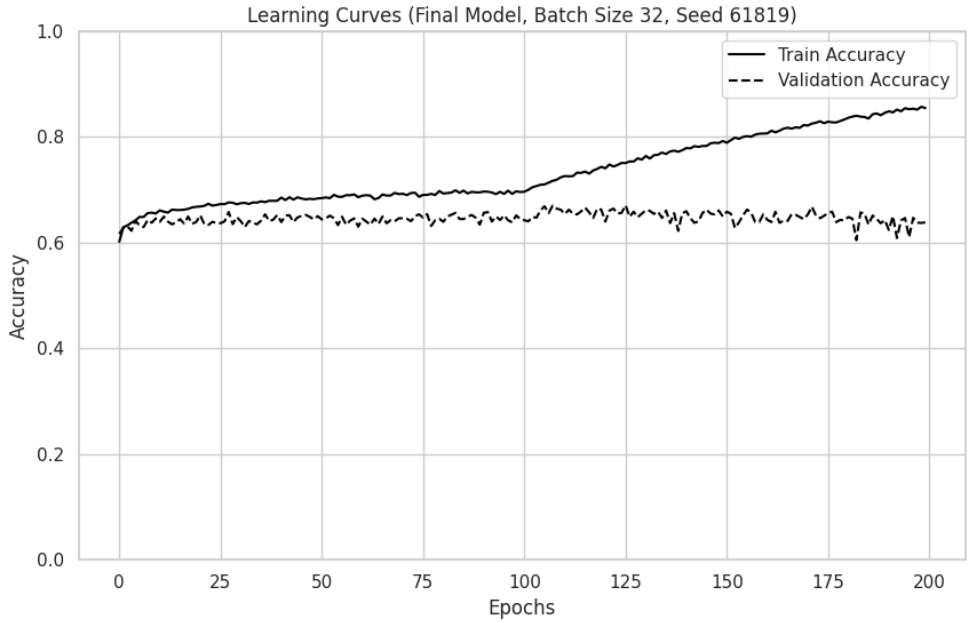


Figure 13: ShuffleNetV2 & AlexNet Learning Curves on CBIS-DDSM Dataset

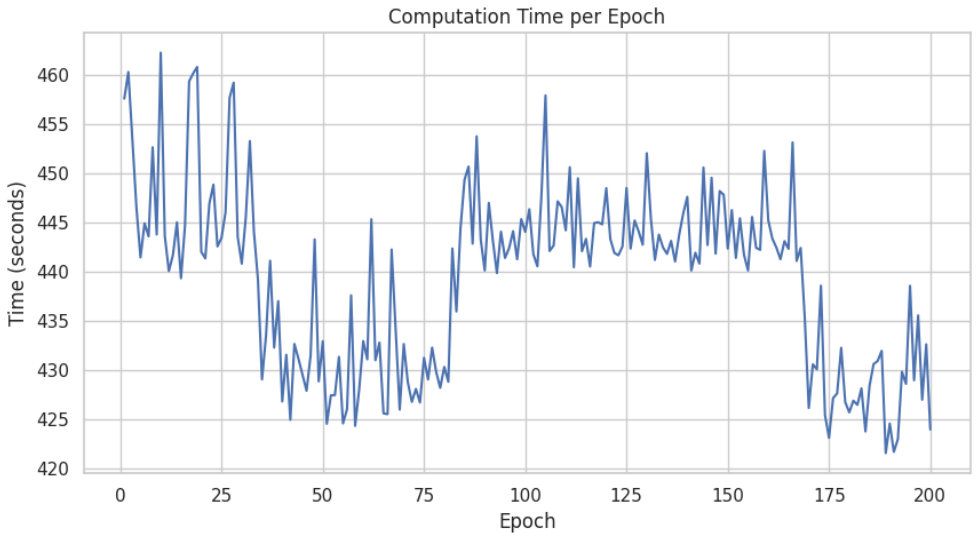


Figure 14: Computation Time in Seconds per Epoch During Training

Figure 14 illustrates the computation time per epoch during the model's training. Notably, the training utilised a Tesla V100 GPU, so the computation times are specific to the Tesla V100, and the performance on other hardware may vary significantly. However, on the Tesla V100, the average training time per epoch was approximately 445 seconds, with the range being approximately 425 to 465 seconds.

The ShuffleNetV2 and AlexNet architecture demonstrated promising generalisability across datasets, achieving a test set accuracy of 66.57% and maintaining a consistent weighted F1-Score of 0.65 on both validation and test sets. However, despite these stable overall metrics, the model exhibited significant shifts in prediction patterns between the validation and test sets, favoring malignant diagnoses in validation but benign diagnoses in testing. The training process on a Tesla V100 GPU with an average epoch time of 445 seconds showed effective initial learning but signs of over-fitting in the second stage. While the model shows potential for cross-dataset application, the inconsistency in prediction patterns and the onset of over-fitting highlight areas for further refinement to enhance its reliability and performance stability across different datasets.

7 Discussion

7.1 Summary of Key Findings

The study presents several findings that contribute to mammogram classification using deep learning techniques. Notably, the ShuffleNetV2 and AlexNet hybrid model outperforms alternative hybrids and vision transformers, as witnessed in Chapter 5 and align with previous findings [5]. The model then demonstrates generalisation capabilities on an unseen dataset in Chapter 6 despite time disallowing a proper optimisation of the model, scoring a test set accuracy of 66.57%. The F1 score is higher than 54% of the studies using radiographers included in a review, with the highest only generating an F1-score of 0.689 compared to our F1-Score of 0.65[115], indicating good performance against radiographers with training and experience. However, the score of 66.57% accuracy on the test set fails to reproduce the claims of 99.17% accuracy when deployed with over-fitting controls as suggested[5], but does outperform similar studies using CBIS-DDSM as a test set[7][8].

7.2 Analysis of Batch Size and Seed Effects

The investigation into batch sizes across the primary and secondary models indicates that batch sizes of 32 are favourable for yielding higher validation accuracies on average (as presented in Table 19 of the appendix) than larger batch sizes of 126 or 252 and smaller batch sizes of 16. Additionally, batch size 32 is the fastest to converge for the primary model, as Table 18 in the appendix presents. Batch size 32's performance may be attributed to the balance between gradient noise and stability. The balance allows for practical exploration of the loss landscape and better generalisation, avoiding the pitfalls of overfitting that can occur with larger batches. This finding fills a hole in the literature regarding batch size effects in mammogram classification, contradicting literature on batch size effects in alternative settings such as on the MNIST handwriting dataset[116] but is supported by image classification with deep networks on ImageNet[117]. Future work should investigate whether the observed effectiveness is influenced by the proportion of the dataset size it represents. In this case, the batch size of 32 was optimal across dataset sizes of 5202 and 28,176.

The investigation into seed initialisation also plays a crucial role in classification performance and as a benchmarker for reproducibility [118]. The approach adopted here is one of interval seeds, and it was found that particular seed values consistently resulted in higher validation accuracies, as identified in Table 7 of the Validation Chapter. The observation highlights seed optimisation's critical nature to enhance reproducibility and classification performance. The slight variance identified for the Primary Model might seem insignificant, but the impact could be more pronounced in a more extensive search space. A conclusive point that could have led to further optimisation in the generalisability testing had the project been able to trial more configurations.

7.3 Loss Function Selection

The performance of different loss functions was a critical aspect of this study. Contrary to expectations, categorical cross-entropy outperformed binary cross-entropy, achieving the highest validation accuracy of 65.33%. This result is exciting given that binary cross-entropy is widely recommended[19] and used in binary classification tasks such as mammogram classification[6][86][83]. The key difference that could explain this result lies in the output neuron structure: categorical cross-entropy uses two softmax-activated output neurons, which may have enabled the model

to differentiate between benign and malignant cases more effectively. This finding suggests that this choice of the loss function can impact model performance, particularly in complex tasks[92], potentially aiding to differentiate subtle in-group differences.

7.4 Optimiser and Learning Rate Performance

Casting aspersions on the prior findings of categorical cross-entropies results, the investigation into optimisers yields conflicting results. In the Primary Model, the study finds that the Adam optimiser surprisingly does not yield the highest validation accuracy. However, when testing for the optimiser in the Secondary Model, the study finds that across more trials and data, the Adam optimiser consistently yields better results, which is consistent with Adam's use case and prevalence in the literature[108][37][6][20][14].

7.5 Performance of Hybrid Models

The hybrid models developed in this study showcased varying performance levels, with the ShuffleNetV2 and AlexNet combination emerging as the most successful, achieving a validation accuracy of 70.12% in the validation phase. This section delves into the architectural features and the underlying reasons behind these hybrid models' varying performances, supported by relevant literature and a discussion on hyperparameter optimisation.

ShuffleNetV2 AlexNet Hybrid
The ShuffleNetV2 and AlexNet hybrid model is the most effective among the tested architectures across the fifty optimisation trials reproducing parent study results[5]. ShuffleNetV2 and AlexNet's success can be attributed to the complementary strengths of its components. ShuffleNetV2 is designed for computational efficiency, leveraging group convolution and channel shuffling to reduce computational costs while maintaining accuracy, making it particularly useful in resource-constrained environments (Zhang et al., 2018). Additionally, the fewer parameters compared to the more extensive architecture of AlexNet enable the model to learn without over-fitting quickly, as seen in Figure 13. On the other hand, AlexNet excels at feature extraction, having proven its effectiveness in large-scale image classification tasks (Krizhevsky et al., 2012). Combining these two architectures and their pre-trained weights allowed the model to fine-tune effectively.

ShuffleNetV2 handles the initial stages of feature extraction, benefiting from its efficiency in handling significant inputs with reduced computational load. AlexNet then refined these features, capitalising on its more profound and complex layers to enhance classification accuracy. The balance between the lightweight, efficient architecture of ShuffleNetV2 and the deep, feature-rich architecture of AlexNet was vital to achieving the highest validation accuracy in this study.

However, this study failed to reproduce the claims from a previous study achieving 99.17% accuracy [5]. From the investigations and comprehensive nature of the documentation of this study, the details undocumented in the previous study, and the concerns of over-fitting identified, this study concludes that the claims are unrepeatable. This study identified items that the prior left out in Table 6, which are believed to have made the difference in testing accuracy's. Specifically problematic, the study states that the data-splitting strategy employs cross-validation but does not identify an isolated test set split. Hence, the weights could have leaked across folds of the training

process if not re-initialised between folds - which the authors should document [5]. However, the difference could be from more extensive augmentation, the number of trials for hyperparameter optimisation, and specific pre-processing parameters.

Alternative Hybrids
The InceptionV3 ResNet18 hybrid model was a novel architectural choice inspired by previous works [7][5][86][20], achieving a validation accuracy of 68.46%. However, despite the strengths identified in the literature review and methodology, the model's performance was slightly lower than that of the ShuffleNetV2 and AlexNet hybrid.

On the other hand, the ShuffleNetV2 and ResNet18 hybrid, which was directly based on the work by Sahu [5], reached a validation accuracy of 67.97%. While this hybrid was not a novel combination, it was expected to perform well due to reported success[5]. Despite this, the model still fell short of the performance of our novel hybrid model and the ShuffleNetV2 and AlexNet hybrid model.

Both hybrids faced challenges related to over-fitting and the need for extensive hyperparameter tuning. With its higher complexity, the InceptionV3 and ResNet18 model may have required more careful balancing of hyperparameters, like dropout rates and regularisation strengths, to avoid over-fitting. Although it used a dropout rate of 0.6, this did not prevent over-fitting, as training accuracy reached 100% without a corresponding increase in validation accuracy, as shown in Figure 15 of the appendix.

The ShuffleNetV2 and ResNet18 hybrid, despite its theoretical and reported advantages, also struggled with over-fitting, as indicated in Figure 16. One explanation could be the vast hyperparameter search space and the limited number of trials (fifty) to explore this space effectively. The optimisation process could have benefited from more sophisticated methods, such as Bayesian Optimisation within the Optuna framework, to identify the optimal hyperparameters quicker [74][111][70].

Hyperparameter Optimisation - Commonalities and Differences
Specific hyperparameters were consistent across the hybrid models, such as using the Adam optimiser and a base batch size of 32 for all models except the ResNet50, MobileNetV2, and InceptionV3 hybrid, which used a batch size of 48. The consistency in optimiser choice reflects Adam's effectiveness in handling sparse gradients, which is common in deep networks with many layers[108]. However, the variation in learning rates, dropout rates, and regularisation strengths across the models suggests that the optimal configuration for these parameters depends highly on the specific architecture. For instance, the lower learning rate in the ShuffleNetV2 and AlexNet hybrid allowed for more gradual convergence. Similarly, the differing dropout rates across models indicate that while dropout is a proper regularisation technique, its optimal rate depends on the network's depth and complexity.

7.6 Ethical Considerations

While the research presents competitive diagnostic accuracy on the test set and other research claims better, there are significant problems including but not limited to accuracy, bias, explainability, accountability, and job displacement[119][120].
While attaining high accuracies on the provided datasets in the public domain is an achievement,

there are still critical issues regarding data representation. For example, our datasets are from Chinese and American citizens[51] [52], which could result in higher misdiagnoses in underrepresented groups. Additionally, the models with high accuracy and generalisability could result in job displacement of healthcare professionals[120].
Explanability is another ethical concern [119]. Convolutional Neural Networks operate in a manner that the chain of reasoning is anonymised in the many layers and neurons[54], which is challenging to trust[55], especially in healthcare [56]. Accountability is also an issue, as technological failures can directly lead to patient harm[121].

7.7 Limitations

One limitation of the pre-processing pipeline may be the absence of adaptive median filtering as is utilised in much of the literature for this task[6][86][88][21]. Another significant limitation throughout the project was the time constraints inherent to a dissertation project. Most notably, this time constraint restricted the number of hyperparameter optimisation trials run for the hybrid models during validation (to fifty) and during testing (to five), and also, two of the vision transformer models were not completed in time for fair reporting of their best results.

Another limitation of the research is the size of the datasets used, as even if combined, the datasets total only 3331 patients[51][52], a tiny fraction of the estimated two million scans performed annually [1]. A further limitation of the study is the need for error analysis implementation. Without the details of incorrectly classified patients, the study cannot manually examine trends in incorrect predictions, for example, due to errors in pre-processing.

8 Conclusion

The thesis objective was to generate several pipelines capable of classifying mammograms using deep-learning techniques based on the literature review. The thesis achieves all the primary, secondary and tertiary objectives outlined in the Introduction. The implementations show promise in aiding diagnostic accuracy and represent an opportunity to reduce misdiagnosed breast cancer cases and consequent mortality with further work.

Initially, datasets from the cancer imaging archive were downloaded, converted to PNGs from DICOM, processed using the Wiener Filter, CLAHE, cropped if CBIS-DDSM, and augmented. Then, a Primary Model was constructed to perform parameter investigations, such as testing different loss functions, optimisers and augmentations, which helped inform the implementations of the hybrid models. Then, the hybrid set of models, inspired by previous works[5][7][6], was crafted - with some novel architectures experimented. These hybrid models combined top-performing models on ImageNet[31] and fine-tuned their weights to classify mammograms. These models were optimised using Optuna and Keras-Tuner to find the best hyperparameters over an extensive search space. However, the number of trials was limited by time, potentially limiting the effectiveness of the hyperparameters.

The paper presents several findings. Most importantly, the paper identifies that the ShuffleNetV2 and AlexNet hybrid is highly generalisable and classifies mammograms the best with a test set accuracy of 66.57%, beating comparable thesis projects[7][8] despite failing to reproduce claims of

99.17% accuracy[5]. Secondary to the failed reproducibility, the paper presents a reproducibility checklist, outlining a start towards better reporting aligning with scientific principles. Further, the study finds the extent to which geometric transformations aid validation accuracy and that categorical cross-entropy instead of binary cross-entropy may aid the classification of binary tasks.

The project has contributed several findings to a critical research field by narrowing down the search space of effective parameters critical to the pipeline which classifies mammograms the most effectively. Future researchers can use our reproducibility checklist and even build on it to prune the search space in this area. With further exploration, these findings could be applied to larger, more diverse datasets to enhance the generalisability of the models. Eventually, integrating these techniques into real-world clinical settings could revolutionise how radiologists approach breast cancer screening, potentially increasing early detection rates and reducing diagnostic errors. Future work should focus on refining these models, exploring recent developments in vision transformers, and addressing potential biases in datasets to create more equitable and effective diagnostic tools. Future research should also strive to build tools that can aid, and not entirely replace, radiographers. For instance, the current recommendation is for double readings of all mammography examinations, which is an economic burden[122][12]. Future work should use artificial intelligence to minimise the need for double-readings, allowing radiographers more time for complex cases. Building on the foundation laid by this thesis, future research could significantly contribute to saving lives and improving patient outcomes in breast cancer care.

9 Future Work

Future work should explore alternative pre-processing techniques like adaptive median filtering to improve image quality and model performance[6][86][**Arooj**] [**Al-Antari2020**]. Investigating new hybrid models with various fine-tuning strategies, such as adaptive fine-tuning [123], could further enhance adaptability and accuracy. Additional hyperparameter optimisation trials and more effective optimisation methods, such as Bayesian optimisation[74][111], are essential for maximising performance and need exploring further.

Completing and optimising vision transformers studied here, as well as exploring new transformers that are outperforming AlexNet on ImageNet[124][125] alongside exploring ensemble methods with sophisticated voting mechanisms, could lead to significant advancements[126].

More extensive and diverse datasets are critical to improving model generalisation and reducing bias, and testing the ShuffleNet and AlexNet hybrid on more data could lead to further improvements. Similarly, future work should test how adding combinations of augmentations improves validation accuracy. Extending training durations to explore grokking effects might uncover new performance optima[127]. Finally, addressing CNNs' "black-box" nature by integrating explainability tools like heat maps and attention mechanisms will enhance transparency, trust, and clinical usability[57].

10 Contributions to Knowledge

The research presents contributions to knowledge across the pipeline for mammogram classification in several areas. First, the paper presents findings of the augmentations that boost classification

accuracy when combined with the original dataset. Second, the paper presents different parameter results for Wiener Filtering and Contrast Limited Adaptive Histogram Equalisation. Third, the paper identifies a batch size of 32 as consistently the highest performing in terms of classification and convergence speed. Fourth, the paper finds that categorical cross-entropy with two softmax output neurons performs better than binary cross-entropy in a binary classification task. Fifth, the paper finds RMSprop converges 4.85% faster with 1.63% higher validation accuracy on average on the CMMD dataset, so it is viable for initial model trialling on smaller datasets. At the same time, Adam performs better on larger CBIS-DDSM dataset with augmentations. Sixth, the research presents an itinerary for researchers and publishers to assess the scientific principle of reproducibility. Seventh, the research presents ShuffleNet and AlexNet, a lightweight unfrozen architecture paired with a deeper frozen architecture that performs better than alternative hybrids. The study finds that adding a third model to the hybrids does not perform as well as two models - though more rigorous investigations are needed. Finally, the paper finds that while vision transformers show promise, the unoptimised versions do not perform as well as the hybrid convolutional neural networks. The study found no improvements over brief hyperparameter optimisation for the vision transformers.

11 References

References

[1] Cancer Research Fund International. *Breast Cancer Statistics 2022*. https://www.who.int/. Accessed: 2024-08-03. 2024.

[2] Cancer Imaging Archive. *Annual Mammogram Statistics 2022*. https://www.cancerimagingarchive.net/. Accessed: 2024-08-03. 2024.

[3] FJHM Van den Biggelaar, PJ Nelemans, and K Flobbe. "Performance of radiographers in mammogram interpretation: a systematic review". In: *The Breast* 17.1 (2008), pp. 85–90.

[4] BreastCancer.org. *Artificial Intelligence in Breast Cancer Screening and Testing*. https://www.breastcancer.org/screening-testing/artificial-intelligence. Accessed: 2024-08-15. 2024.

[5] A. Sahu, P.K. Das, and S. Meher. "High accuracy hybrid CNN classifiers for breast cancer detection using mammogram and ultrasound datasets". In: *Biomedical Signal Processing and Control* 80 (2023), p. 104292.

[6] Y. Thangavel et al. "Revolutionizing breast cancer diagnosis with a comprehensive approach using digital mammogram-based feature extraction and selection for early-stage identification". In: *Biomedical Signal Processing and Control* 94 (2024), p. 106268. DOI: 10.1016/j.bspc.2024.106268.

[7] Adam Jaamour et al. "A divide and conquer approach to maximise deep learning mammography classification accuracies". In: *PLOS ONE* 18.5 (2023), e0280841. DOI: 10.1371/journal.pone.0280841.

[8] Rhona McCracken. *CS5199 Breast Cancer Detection Project*. Accessed: 2024-08-04. 2024. URL: https://github.com/RAMcCracken/CS5199_Breast_Cancer_Detection_Project.

[9] E. M. Rosen et al. "BRCA1 gene in breast cancer". In: *Journal of Cellular Physiology* 196.1 (2003), pp. 19–41. DOI: 10.1002/jcp.10206.

[10] National Institute for Health Research. *Breast Cancer Screening: Reducing Mortality Rates*. https://www.nihr.ac.uk/. Accessed: 2024-08-03. 2023.

[11] Lazaros Tsochatzidis, Lena Costaridou, and Ioannis Pratikakis. "Deep learning for breast cancer diagnosis from mammograms—a comparative study". In: *Journal of Imaging* 5.3 (2019), p. 37.

[12] Catherine Chilute Chilanga, Hilde Merete Olerud, and Kristin Bakke Lysdahl. "Radiographers' actions and challenges when confronted with inappropriate radiology referrals". In: *European Radiology* 32.6 (2022), pp. 4210–4217.

[13] A. Krizhevsky, I. Sutskever, and G.E. Hinton. "ImageNet classification with deep convolutional neural networks". In: *Advances in Neural Information Processing Systems*. Vol. 25. 2012, pp. 1097–1105.

[14] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[15] Craig Macfadyen, Ajay Duraiswamy, and David Harris-Birtill. "Classification of hyper-scale multimodal imaging datasets". In: *PLOS Digital Health* 2.12 (2023), e0000191. DOI: 10.1371/journal.pdig.0000191.

[16] S. Reddy, J. Fox, and M.P. Purohit. "Artificial intelligence-enabled healthcare delivery". In: *Journal of the Royal Society of Medicine* 112.1 (2019), pp. 22–28.

[17] Adrian P. Brady. "Error and discrepancy in radiology: inevitable or avoidable?" In: *Insights into Imaging* 8.1 (2017), pp. 171–182. DOI: 10.1007/s13244-016-0534-1. URL: https://doi.org/10.1007/s13244-016-0534-1.

[18] G. Litjens et al. "A survey on deep learning in medical image analysis". In: *Medical Image Analysis* 42 (2017), pp. 60–88.

[19] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[20] K.K. Dewangan et al. "Breast cancer diagnosis in an early stage using novel deep learning with hybrid optimization technique". In: *Multimedia Tools and Applications* 81.10 (2022), pp. 13935–13960.

[21] M.A. Al-Antari, S.M. Han, and T.S. Kim. "Evaluation of deep learning detection and classification towards computer-aided diagnosis of breast lesions in digital X-ray mammograms". In: *Computer Methods and Programs in Biomedicine* 196 (2020), p. 105584.

[22] Y. LeCun, Y. Bengio, and G. Hinton. "Deep learning". In: *Nature* 521.7553 (2015), pp. 436–444.

[23] D.R. Nayak et al. "Pathological brain detection using Improved Jaya algorithm (IJaya) and orthogonal ripplet-II transform (O-DR2T)". In: *Journal of King Saud University-Computer and Information Sciences* 34.1 (2018), pp. 74–83.

[24] N. Muduli, S. Kumar, and V. Bhateja. "Fast discrete curvelet transform based breast cancer detection using modified particle swarm optimisation and extreme learning machine". In: *Journal of Ambient Intelligence and Humanized Computing* 12 (2021), pp. 11617–11630.

[25] V.K. Singh et al. "Conditional generative adversarial and convolutional networks for X-ray breast mass segmentation and shape classification". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2018: 21st International Conference*. 2018, pp. 833–840.

[26] A. Vaswani et al. "Attention is all you need". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017, pp. 5998–6008.

[27] Ian Goodfellow et al. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems*. Vol. 27. 2014. URL: https://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

[28] Y. LeCun et al. "Backpropagation applied to handwritten zip code recognition". In: *Neural Computation* 1.4 (1989), pp. 541–551.

[29] Vinod Nair and Geoffrey E. Hinton. "Rectified Linear Units Improve Restricted Boltzmann Machines". In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. 2010, pp. 807–814. URL: https://icml.cc/Conferences/2010/papers/432.pdf.

[30] N. Srivastava et al. "Dropout: a simple way to prevent neural networks from overfitting". In: *Journal of Machine Learning Research* 15.1 (2014), pp. 1929–1958.

[31] Jia Deng et al. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[32] D. Ciresan et al. "Mitosis detection in breast cancer histology images with deep neural networks". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013*. 2012, pp. 411–418.

[33] O. Ronneberger, P. Fischer, and T. Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*. 2015, pp. 234–241.

[34] K. Cho, S. Kumar, and C. Matsoukas. "Accelerating deep learning workloads on NVIDIA GPUs: a deep learning framework perspective". In: *IEEE Transactions on Computers* 70.6 (2020), pp. 970–981.

[35] K. Clark et al. "The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository". In: *Journal of Digital Imaging* 26.6 (2013), pp. 1045–1057.

[36] K. Simonyan and A. Zisserman. *Very deep convolutional networks for large-scale image recognition*. arXiv preprint arXiv:1409.1556. 2014.

[37] K. He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 770–778.

[38] D. Shen, G. Wu, and H.I. Suk. "Deep learning in medical image analysis". In: *Annual Review of Biomedical Engineering* 19 (2017), pp. 221–248.

[39] M.D. Zeiler and R. Fergus. "Visualizing and understanding convolutional networks". In: *European Conference on Computer Vision (ECCV)*. 2014.

[40] N. Shahriar. *What is Convolutional Neural Network (CNN)? [Deep Learning]*. Medium. Accessed 3 August 2024. 2020. URL: https://nafizshahriar.medium.com/what-is-convolutional-neural-network-cnn-deep-learning-b3921bdd82d5.

[41] Y. LeCun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[42] Andre Esteva et al. "Dermatologist-level classification of skin cancer with deep neural networks". In: *Nature* 542.7639 (2017), pp. 115–118. DOI: 10.1038/nature21056. URL: https://www.nature.com/articles/nature21056.

[43] P. Ghosal, R.K. Dwivedi, and A.S. Ashour. "Noise-robust feature extraction and classification framework for automated pulmonary abnormality detection in chest radiographs". In: *Computer Methods and Programs in Biomedicine* 181 (2019), p. 104830.

[44] Ahmed Hosny et al. "Artificial intelligence in radiology". In: *Nature Reviews Cancer* 18.8 (2018), pp. 500–510. DOI: 10.1038/s41568-018-0016-5. URL: https://www.nature.com/articles/s41568-018-0016-5.

[45] X. Zhao et al. "A review of convolutional neural networks in computer vision". In: *Artificial Intelligence Review* 57.4 (2024), p. 99.

[46] M. Buda, A. Maki, and M.A. Mazurowski. "A systematic study of the class imbalance problem in convolutional neural networks". In: *Neural Networks* 106 (2018), pp. 249–259.

[47] Rikiya Yamashita et al. "Convolutional neural networks: an overview and application in radiology". In: *Insights into Imaging* 9.4 (2018), pp. 611–629. DOI: 10.1007/s13244-018-0639-9. URL: https://link.springer.com/article/10.1007/s13244-018-0639-9.

[48] Gavin C. Cawley and Nicola L. C. Talbot. "On over-fitting in model selection and subsequent selection bias in performance evaluation". In: *Journal of Machine Learning Research* 11 (2010), pp. 2079–2107. URL: http://jmlr.org/papers/v11/cawley10a.html.

[49] L. Perez and J. Wang. *The effectiveness of data augmentation in image classification using deep learning*. arXiv preprint arXiv:1712.04621. 2017.

[50] F. Garcea et al. "Data augmentation for medical imaging: A systematic literature review". In: *Computers in Biology and Medicine* 152 (2023), p. 106391.

[51] C. Cui et al. *The Chinese Mammography Database (CMMD): An online mammography database with biopsy confirmed types for machine diagnosis of breast*. The Cancer Imaging Archive. 2021. URL: https://doi.org/10.7937/tcia.eqde-4b16.

[52] R.S. Lee et al. "A curated mammography data set for use in computer-aided detection and diagnosis research". In: *Scientific Data* 4.1 (2017). DOI: 10.1038/sdata.2017.177. URL: https://doi.org/10.1038/sdata.2017.177.

[53] S.J. Pan and Q. Yang. "A survey on transfer learning". In: *IEEE Transactions on Knowledge and Data Engineering* 22.10 (2010), pp. 1345–1359.

[54] Alexander S. Lundervold and Arvid Lundervold. "An overview of deep learning in medical imaging focusing on MRI". In: *Zeitschrift für Medizinische Physik* 29.2 (2019), pp. 102–127. DOI: 10.1016/j.zemedi.2018.11.002.

[55] M.T. Ribeiro, S. Singh, and C. Guestrin. ""Why should I trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778.

[56] R. Caruana et al. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015, pp. 1721–1730. DOI: 10.1145/2783258.2788613.

[57] R.R. Selvaraju et al. "Grad-CAM: Visual explanations from deep networks via gradient-based localisation". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 618–626. DOI: 10.1109/ICCV.2017.74.

[58] P. Lambin et al. "Radiomics: extracting more information from medical images using advanced feature analysis". In: *European Journal of Cancer* 48.4 (2012), pp. 441–446.

[59] G. James et al. *An Introduction to Statistical Learning: With Applications in R*. New York: Springer, 2013.

[60] K. Zuiderveld. "Contrast limited adaptive histogram equalisation". In: *Graphics Gems IV*. Ed. by P. Heckbert. San Diego: Academic Press, 1994, pp. 474–485.

[61] I.T. Jolliffe. *Principal Component Analysis*. 2nd. New York: Springer, 2002.

[62] R.A. Fisher. "The use of multiple measurements in taxonomic problems". In: *Annals of Eugenics* 7.2 (1936), pp. 179–188.

[63] Emmanuel J. Candes and David L. Donoho. "Continuous Curvelet Transform: I. Resolution of the Wavefront Set". In: *Applied and Computational Harmonic Analysis* 19.2 (2005), pp. 162–197. DOI: 10.1016/j.acha.2005.02.005.

[64] James Kennedy and Russell Eberhart. "Particle Swarm Optimization". In: *Proceedings of ICNN'95 - International Conference on Neural Networks*. IEEE. 1995, pp. 1942–1948. DOI: 10.1109/ICNN.1995.488968.

[65] G.B. Huang, Q.Y. Zhu, and C.K. Siew. "Extreme learning machine: theory and applications". In: *Neurocomputing* 70.1-3 (2006), pp. 489–501.

[66] Z. Khandezamin, M. Naderan, and M.J. Rashti. "Detection and classification of breast cancer using logistic regression feature selection and GMDH classifier". In: *Journal of Biomedical Informatics* 111 (2020), p. 103591.

[67] D.R. Cox. "The regression analysis of binary sequences". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 20.2 (1958), pp. 215–232.

[68] A.G. Ivakhnenko. "The group method of data handling—A rival of the method of stochastic approximation". In: *Soviet Automatic Control* 13.3 (1968), pp. 43–56.

[69] P. Domingos. "A few useful things to know about machine learning". In: *Communications of the ACM* 55.10 (2012), pp. 78–87.

[70] E. Akkur, F. Turk, and O. Erogul. "Breast Cancer Diagnosis Using Feature Selection Approaches and Bayesian Optimization". In: *Computer Systems Science & Engineering* 45.2 (2023).

[71] K. Kira and L.A. Rendell. "The feature selection problem: Traditional methods and a new algorithm". In: *Proceedings of the Tenth National Conference on Artificial Intelligence*. 1992, pp. 129–134.

[72] R. Tibshirani. "Regression shrinkage and selection via the Lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[73] A.W. Whitney. "A direct method of nonparametric measurement selection". In: *IEEE Transactions on Computers* 100.9 (1971), pp. 1100–1103.

[74] J. Snoek, H. Larochelle, and R.P. Adams. "Practical Bayesian optimization of machine learning algorithms". In: *Advances in Neural Information Processing Systems*. Vol. 25. 2012.

[75] I. Guyon and A. Elisseeff. "An introduction to variable and feature selection". In: *Journal of Machine Learning Research* 3 (2003), pp. 1157–1182.

[76] D. Delen, G. Walker, and A. Kadam. "Predicting breast cancer survivability: a comparison of three data mining methods". In: *Artificial Intelligence in Medicine* 34.2 (2005), pp. 113–127.

[77] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. Wiley, 1973.

[78] P. Domingos and M. Pazzani. "On the optimality of the simple Bayesian classifier under zero-one loss". In: *Machine Learning* 29.2-3 (1997), pp. 103–130.

[79] D.J. Hand and K. Yu. "Idiot's Bayes—not so stupid after all?" In: *International Statistical Review* 69.3 (2001), pp. 385–398.

[80] Murat Karabatak. "A new feature selection method based on association rules for breast cancer detection". In: *Journal of Medical Systems* 39.2 (2015), p. 21. DOI: 10.1007/s10916-015-0191-5.

[81] W.S. McCulloch and W. Pitts. "A logical calculus of the ideas immanent in nervous activity". In: *The Bulletin of Mathematical Biophysics* 5.4 (1943), pp. 115–133.

[82] F. Rosenblatt. "The perceptron: A probabilistic model for information storage and organization in the brain". In: *Psychological Review* 65.6 (1958), pp. 386–408.

[83] M.H. Alshayeji, H. Ellethy, and R. Gupta. "Computer-aided detection of breast cancer on the Wisconsin dataset: An artificial neural networks approach". In: *Biomedical Signal Processing and Control* 71 (2022), p. 103141.

[84] C.M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[85] T. H. Chen and A. R. Ramli. "Contrast enhancement using recursive mean-separate histogram equalization for scalable brightness preservation". In: *IEEE Transactions on Consumer Electronics* 49.4 (2004), pp. 1301–1309. DOI: 10.1109/TCE.2003.1261234.

[86] S.S. Chakravarthy and H. Rajaguru. "Automatic detection and classification of mammograms using improved extreme learning machine with deep learning". In: *IRBM* 43.1 (2022), pp. 49–61.

[87] Ali Askarzadeh. "A novel metaheuristic method for solving constrained engineering optimization problems: Crow search algorithm". In: *Computers & Structures* 169 (2016), pp. 1–12. DOI: 10.1016/j.compstruc.2016.03.001.

[88] S. Arooj et al. "Breast cancer detection and classification empowered with transfer learning". In: *Frontiers in Public Health* 10 (2022), p. 924432.

[89] Mark Sandler et al. "Mobilenetv2: Inverted residuals and linear bottlenecks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4510–4520.

[90] X. Zhang et al. "ShuffleNet: An extremely efficient convolutional neural network for mobile devices". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pp. 6848–6856.

[91] A. Krizhevsky, I. Sutskever, and G.E. Hinton. "ImageNet classification with deep convolutional neural networks". In: *Advances in Neural Information Processing Systems*. Vol. 25. 2012, pp. 1097–1105.

[92] C. Szegedy et al. "Going deeper with convolutions". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 1–9.

[93] Q. Zhang and Y. Wu. "Adaptive fuzzy filter for image enhancement". In: *IEEE Transactions on Image Processing* 17.5 (2008), pp. 664–674.

[94] D.L. Donoho. "De-noising by soft-thresholding". In: *IEEE Transactions on Information Theory* 41.3 (1995), pp. 613–627.

[95] S.M. Pizer et al. "Adaptive histogram equalization and its variations". In: *Computer Vision, Graphics, and Image Processing* 39.3 (1987), pp. 355–368.

[96] N. Wiener. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. MIT Press, 1949.

[97] T.S. Huang, G.J. Yang, and G.Y. Tang. "A fast two-dimensional median filtering algorithm". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 27.1 (1979), pp. 13–18.

[98] N. Otsu. "A threshold selection method from gray-level histograms". In: *IEEE Transactions on Systems, Man, and Cybernetics* 9.1 (1979), pp. 62–66.

[99] J. MacQueen. "Some methods for classification and analysis of multivariate observations". In: *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 1. 14. 1967, pp. 281–297.

[100] J. Serra. *Image Analysis and Mathematical Morphology*. Academic Press, 1982.

[101] R. Adams and L. Bischof. "Seeded region growing". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16.6 (1994), pp. 641–647.

[102] R. Kohavi. "A study of cross-validation and bootstrap for accuracy estimation and model selection". In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. Vol. 2. 1995, pp. 1137–1143.

[103] Jakub Nalepa, Michal Kawulok, and Piotr Antosz. "Data Augmentation for Brain-Tumor Segmentation: A Review". In: *Frontiers in Computational Neuroscience* 13 (2019), p. 83. DOI: 10.3389/fncom.2019.00083. URL: https://www.frontiersin.org/articles/10.3389/fncom.2019.00083/full.

[104] D.A. Ragab et al. "A framework for breast cancer classification using multi-DCNNs". In: *Computers in Biology and Medicine* 131 (2021), p. 104245.

[105] Ningning Ma et al. "Shufflenet v2: Practical guidelines for efficient cnn architecture design". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 116–131.

[106] Solomon Kullback and Richard A. Leibler. "On Information and Sufficiency". In: *The Annals of Mathematical Statistics* 22.1 (1951), pp. 79–86. DOI: 10.1214/aoms/1177729694.

[107] Herbert Robbins and Sutton Monro. "A Stochastic Approximation Method". In: *The Annals of Mathematical Statistics* 22.3 (1951), pp. 400–407.

[108] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization". In: *arXiv preprint arXiv:1412.6980* (2014). URL: https://arxiv.org/abs/1412.6980.

[109] Marco Cantone et al. "Convolutional networks and transformers for mammography classification: an experimental study". In: *Sensors* 23.3 (2023), p. 1229.

[110] Tom O'Malley et al. *Keras Tuner*. https://github.com/keras-team/keras-tuner. Accessed: 2024-08-15. 2019.

[111] Takuya Akiba et al. "Optuna: A next-generation hyperparameter optimization framework". In: *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2019, pp. 2623–2631.

[112] James Bergstra et al. "Algorithms for hyper-parameter optimization". In: *Advances in neural information processing systems* 24 (2011).

[113] James Bergstra and Yoshua Bengio. "Random search for hyper-parameter optimization." In: *Journal of machine learning research* 13.2 (2012).

[114] Lisha Li et al. "Hyperband: A novel bandit-based approach to hyperparameter optimization". In: *Journal of Machine Learning Research* 18.185 (2018), pp. 1–52.

[115] S Moran and H Warren-Forward. "The diagnostic accuracy of radiographers assessing screening mammograms: A systematic review". In: *Radiography* 22.2 (2016), pp. 137–146.

[116] Pavlo M Radiuk. "Impact of training set batch size on the performance of convolutional neural networks for diverse datasets". In: (2017).

[117] Dominic Masters and Carlo Luschi. "Revisiting small batch training for deep neural networks". In: *arXiv preprint arXiv:1804.07612* (2018).

[118] Liam Li and Ameet Talwalkar. "Random search and reproducibility for neural architecture search". In: *Uncertainty in artificial intelligence*. PMLR. 2020, pp. 367–377.

[119] Jessica Morley et al. “The ethics of AI in health care: a mapping review”. In: *Social Science & Medicine* 260 (2020), p. 113172.

[120] Fan Li, Nick Ruijs, and Yuan Lu. “Ethics & AI: A systematic review on ethical concerns and related strategies for designing with AI in healthcare”. In: *Ai* 4.1 (2022), pp. 28–53.

[121] Dean F Sittig and Hardeep Singh. “A new socio-technical model for studying health information technology in complex adaptive healthcare systems”. In: *Cognitive Informatics for Biomedicine: Human Computer Interaction in Healthcare* (2015), pp. 59–80.

[122] My von Euler-Chelpin et al. “Screening mammography: benefit of double reading by breast density”. In: *Breast cancer research and treatment* 171 (2018), pp. 767–776.

[123] Yunhui Guo et al. “Spottune: transfer learning through adaptive fine-tuning”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 4805–4814.

[124] Ze Liu et al. “Swin transformer: Hierarchical vision transformer using shifted windows”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 10012–10022.

[125] Zihang Dai et al. “Coatnet: Marrying convolution and attention for all data sizes”. In: *Advances in neural information processing systems* 34 (2021), pp. 3965–3977.

[126] Mitchell Wortsman et al. “Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time”. In: *International conference on machine learning*. PMLR. 2022, pp. 23965–23998.

[127] Alethea Power et al. “Grokking: Generalization beyond overfitting on small algorithmic datasets”. In: *arXiv preprint arXiv:2201.02177* (2022).

12 Ethics Approval

University of St Andrews | FOUNDED 1413

School of Computer Science Ethics Committee

24 May 2024

Dear Brandon,

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:	CS17873	**Approved on:**	24.05.24	**Approval Expiry:**	24.05.29
Project Title:	Leveraging Deep Learning in Medical Imaging for Mammary Carcinoma				
Researcher(s):	Brandon Linnett				
Supervisor(s):	Dr 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

School of Computer Science Ethics Committee
Dr Olexandr Konovalov/Convenor, Jack Cole Building, North Haugh, St Andrews, Fife, KY16 9SX
Telephone: 01334 463273 Email: ethics-cs@st-andrews.ac.uk

A Appendix

Performance Metrics for Batch Size and Seed

Batch Size	Seed	Fold	Training Time (seconds)	Best Validation Accuracy (%)	Precision (Benign)	Precision (Malignant)	Recall (Benign)	Recall (Malignant)	F1-score (Benign)	F1-score (Malignant)
16	0	5	2917.75	63.87	0.52	0.51	0.41	0.62	0.46	0.56
16	2000	5	2700.79	64.58	0.46	0.52	0.28	0.70	0.35	0.59
16	4000	5	3390.02	64.75	0.44	0.53	0.28	0.69	0.35	0.60
16	6000	5	3395.95	63.49	0.47	0.52	0.48	0.50	0.48	0.51
16	8000	5	2909.90	64.58	0.50	0.54	0.42	0.62	0.46	0.58
16	10000	5	3355.81	63.28	0.51	0.50	0.31	0.70	0.38	0.58
32	0	5	3064.67	64.62	0.49	0.53	0.32	0.70	0.39	0.60
32	4000	5	3079.33	62.01	0.51	0.50	0.43	0.59	0.47	0.54
32	6000	5	2661.67	65.22	0.45	0.51	0.35	0.60	0.39	0.55
32	8000	5	2671.92	63.51	0.46	0.52	0.29	0.69	0.36	0.59
32	10000	5	2438.13	64.01	0.49	0.53	0.32	0.69	0.39	0.60
64	0	5	2514.70	62.60	0.49	0.52	0.43	0.58	0.46	0.55
64	2000	5	3125.16	62.25	0.46	0.51	0.44	0.53	0.45	0.52
64	4000	5	3126.34	65.69	0.46	0.51	0.35	0.62	0.40	0.56
64	6000	5	3123.62	65.04	0.48	0.51	0.42	0.57	0.45	0.54
64	8000	5	2717.61	64.34	0.44	0.51	0.33	0.62	0.38	0.56
64	10000	5	3126.53	64.65	0.47	0.52	0.46	0.54	0.47	0.53
128	0	5	3032.07	62.21	0.48	0.51	0.43	0.55	0.45	0.53
128	2000	5	3017.72	62.57	0.48	0.53	0.58	0.44	0.53	0.48
128	4000	5	3019.42	60.45	0.48	0.53	0.53	0.48	0.50	0.50
128	6000	5	2425.16	63.02	0.48	0.53	0.23	0.78	0.31	0.63
128	8000	5	3018.57	64.36	0.48	0.51	0.41	0.59	0.45	0.55
128	10000	5	3026.75	60.35	0.48	0.50	0.46	0.52	0.47	0.51
256	0	5	2899.81	62.01	0.47	0.51	0.35	0.63	0.40	0.56
256	2000	5	2706.67	62.60	0.49	0.52	0.49	0.51	0.49	0.52
256	4000	5	2898.56	61.43	0.48	0.50	0.51	0.47	0.49	0.49
256	6000	5	2899.29	60.94	0.49	0.52	0.37	0.63	0.42	0.57
256	8000	5	2898.43	59.86	0.51	0.50	0.44	0.58	0.47	0.54
256	10000	5	2897.90	62.45	0.47	0.51	0.53	0.44	0.50	0.48

Table 17: Performance metrics for various batch sizes and seeds

Batch Size	Validation Accuracy Std Dev
16	0.585
32	1.202
64	1.399
128	1.563
256	0.961

Table 18: Standard deviation of validation accuracy for various batch sizes

Table 19: Best validation accuracy for each batch size with corresponding seed and other metrics

Batch Size	Seed	Fold	Best Validation Accuracy (%)	F1-score (Benign)	F1-score (Malignant)
16	4000	5	64.75	0.35	0.60
32	6000	5	65.22	0.39	0.55
64	4000	5	65.69	0.40	0.56
128	8000	5	64.36	0.45	0.55
256	2000	5	62.60	0.49	0.52

Table 20: Index of Training Times by Batch Size (normalised to Batch Size 16)

Batch Size	**Index**
16	1.000
32	0.894
64	0.950
128	0.939
256	0.921

Table 21: Average Validation Accuracy by Batch Size

Batch Size	**Average Validation Accuracy (%)**
16	64.09
32	63.87
64	64.10
128	62.16
256	61.55

Performance Metrics for Loss Functions
Primary Model Optimiser Results

Table 22: Model Training and Validation Results for Loss Function

Loss Function	Seed	Fold	Learning Rate	Training Time (seconds)	Validation Loss	Validation Accuracy	Best Validation Accuracy
binary_crossentropy	0	5	0.001	3112.80	0.8373	64.35%	64.35%
binary_crossentropy	2000	5	0.001	3343.12	1.1071	61.27%	64.35%
binary_crossentropy	4000	5	0.001	3555.36	1.5964	61.17%	64.55%
binary_crossentropy	6000	5	0.001	2640.94	0.7040	62.46%	64.84%
binary_crossentropy	8000	5	0.001	3091.25	0.8382	62.96%	64.84%
binary_crossentropy	10000	5	0.001	2646.36	0.6634	63.56%	64.84%
categorical_crossentropy	0	5	0.001	2873.14	0.6828	63.46%	65.33%
categorical_crossentropy	2000	5	0.001	3816.65	2.1912	61.67%	65.33%
categorical_crossentropy	4000	5	0.001	4223.96	2.6118	58.89%	65.33%
categorical_crossentropy	6000	5	0.001	3552.91	1.5154	61.77%	65.33%
categorical_crossentropy	8000	5	0.001	2639.49	0.6675	62.16%	65.33%
categorical_crossentropy	10000	5	0.001	3079.74	0.8136	65.14%	65.33%
kld	0	5	0.001	4676.01	8.7233	45.88%	64.84%
kld	2000	5	0.001	4702.29	8.3392	48.26%	64.84%
kld	4000	5	0.001	4645.29	7.6349	52.63%	64.84%
kld	6000	5	0.001	4670.51	8.4032	47.86%	64.84%
kld	8000	5	0.001	4638.91	8.3872	47.96%	64.84%
kld	10000	5	0.001	4705.75	7.9070	50.94%	64.84%
poisson	0	5	0.001	4014.81	4.3815	51.84%	54.30%
poisson	2000	5	0.001	2899.21	4.1414	54.82%	54.82%
poisson	4000	5	0.001	2653.00	4.1894	54.22%	54.82%
poisson	6000	5	0.001	3837.73	4.3174	52.63%	54.82%
poisson	8000	5	0.001	4331.55	4.4455	51.04%	55.31%
poisson	10000	5	0.001	2881.28	4.2618	53.33%	55.31%
sparse_categorical_crossentropy	0	5	0.001	2894.53	0.7249	63.65%	63.65%
sparse_categorical_crossentropy	2000	5	0.001	3127.39	1.0015	59.19%	63.65%
sparse_categorical_crossentropy	4000	5	0.001	2906.00	0.6934	60.18%	63.65%
sparse_categorical_crossentropy	6000	5	0.001	2699.32	0.6936	64.85%	64.85%
sparse_categorical_crossentropy	8000	5	0.001	3154.41	0.8536	62.86%	64.85%
sparse_categorical_crossentropy	10000	5	0.001	2924.55	0.6709	61.77%	64.85%

Table 23: Best Validation Accuracy for Each Loss Function

Loss Function	Seed	Fold	Learning Rate	Training Time (seconds)	Validation Loss	Best Validation Accuracy
binary_crossentropy	6000	5	0.001	2640.94	0.7040	64.84%
categorical_crossentropy	0	5	0.001	2873.14	0.6828	65.33%
kld	4000	5	0.001	4645.29	7.6349	52.63%
poisson	8000	5	0.001	4331.55	4.4455	55.31%
sparse_categorical_crossentropy	6000	5	0.001	2699.32	0.6936	64.85%

Average Performance Metrics - Original + Variations

Validation Loss	Validation Accuracy	Best Validation Accuracy	Benign Precision	Benign Recall	Benign F1-Score	Benign Support	Malignant Precision	Malignant Recall	Malignant F1-Score	Malignant Support
0.6930	51.51	51.51	0.5315	0.5003	0.4990	223	0.3759	0.5063	0.4191	159

Hybrid Models Supplementary Figures and Data

Table 24: Average Validation Accuracy for Different Loss Functions

Loss Function	**Average Validation Accuracy (%)**
binary_crossentropy	62.63
categorical_crossentropy	62.18
kld	48.92
poisson	52.98
sparse_categorical_crossentropy	62.08

Table 25: Adam Optimiser Validation Results

Optimiser	Learning Rate	Seed	Fold	Training Time (seconds)	Validation Loss	Validation Accuracy (%)
Adam	0.001	0	5	2877.99	0.7226	62.30
Adam	0.0001	0	5	4117.02	0.9559	63.91
Adam	0.001	2000	5	2670.46	0.6646	64.62
Adam	0.0001	2000	5	3480.30	1.0266	62.10
Adam	0.001	4000	5	4337.83	2.6133	61.79
Adam	0.0001	4000	5	4329.58	1.1955	63.10
Adam	0.001	6000	5	3495.66	1.3505	62.90
Adam	0.0001	6000	5	3091.10	0.7953	62.10
Adam	0.001	8000	5	3076.37	1.0570	60.48
Adam	0.0001	8000	5	4319.08	1.0706	61.39
Adam	0.001	10000	5	2455.45	0.6737	60.48
Adam	0.0001	10000	5	4329.83	1.2748	63.00

Table 26: RMSprop Validation Results

Optimiser	Learning Rate	Seed	Fold	Training Time (seconds)	Validation Loss	Validation Accuracy (%)
RMSprop	0.001	0	5	2568.09	0.6924	62.46
RMSprop	0.0001	0	5	4203.07	1.9253	63.37
RMSprop	0.001	2000	5	2564.58	0.7002	63.67
RMSprop	0.0001	2000	5	4201.28	1.9089	63.67
RMSprop	0.001	4000	5	2568.97	0.7137	64.28
RMSprop	0.0001	4000	5	4197.40	2.2524	59.94
RMSprop	0.001	6000	5	2978.78	1.2188	64.08
RMSprop	0.0001	6000	5	4201.45	2.0556	62.87
RMSprop	0.001	8000	5	2980.37	1.0447	65.69
RMSprop	0.0001	8000	5	2972.12	0.9626	64.48
RMSprop	0.001	10000	5	3590.20	1.9997	67.20
RMSprop	0.0001	10000	5	3584.87	1.2785	65.99

Table 27: SGD Validation Results

Optimizer	Learning Rate	Seed	Fold	Training Time (seconds)	Validation Loss	Validation Accuracy (%)
SGD	0.001	0	5	4137.11	0.8704	64.88
SGD	0.001	2000	5	4144.87	0.9706	64.08
SGD	0.0001	2000	5	4042.46	0.7810	59.74
SGD	0.001	4000	5	4136.66	0.7346	62.97
SGD	0.0001	4000	5	3652.42	0.7377	60.24
SGD	0.001	6000	5	4131.56	0.7669	64.88
SGD	0.0001	6000	5	4044.84	0.7509	59.54
SGD	0.001	8000	5	4132.84	0.9272	62.56
SGD	0.0001	8000	5	4058.97	0.7829	58.02
SGD	0.001	10000	5	4122.52	0.8421	65.29
SGD	0.0001	10000	5	2924.91	0.7637	59.74

Table 28: Classification Report for RMSprop Optimiser

Learning Rate	**Seed**	**Class**	**Precision**	**Recall**	**F1-score**	**Support**
0.001	0	Benign	0.47	0.38	0.42	498
0.001	0	Malignant	0.50	0.59	0.54	526
0.0001	0	Benign	0.51	0.56	0.54	498
0.0001	0	Malignant	0.54	0.49	0.52	526
0.001	2000	Benign	0.47	0.27	0.35	486
0.001	2000	Malignant	0.52	0.73	0.61	537
0.0001	2000	Benign	0.51	0.50	0.51	498
0.0001	2000	Malignant	0.54	0.55	0.54	526
0.001	4000	Benign	0.50	0.46	0.48	486
0.001	4000	Malignant	0.54	0.59	0.56	537
0.0001	4000	Benign	0.51	0.45	0.48	498
0.0001	4000	Malignant	0.53	0.60	0.57	526
0.001	6000	Benign	0.47	0.49	0.48	486
0.001	6000	Malignant	0.52	0.49	0.50	537
0.0001	6000	Benign	0.51	0.49	0.50	498
0.0001	6000	Malignant	0.53	0.54	0.54	526
0.001	8000	Benign	0.48	0.49	0.48	486
0.001	8000	Malignant	0.53	0.52	0.52	537
0.0001	8000	Benign	0.47	0.50	0.49	486
0.0001	8000	Malignant	0.52	0.50	0.51	537
0.001	10000	Benign	0.48	0.48	0.48	486
0.001	10000	Malignant	0.53	0.53	0.53	537
0.0001	10000	Benign	0.48	0.45	0.46	486
0.0001	10000	Malignant	0.53	0.56	0.54	537

Batch Size	Seed	Validation Loss	Validation Accuracy	Best Validation Accuracy	Benign Precision	Benign Recall	Benign F1-Score	Benign Support	Malignant Precision	Malignant Recall	Malignant F1-Score	Malignant Support
64	909	0.6899	53.898	53.898	0.5658	0.7765	0.6429	322	0.4441	0.2249	0.2521	248

Table 29: Average Table - translated 300

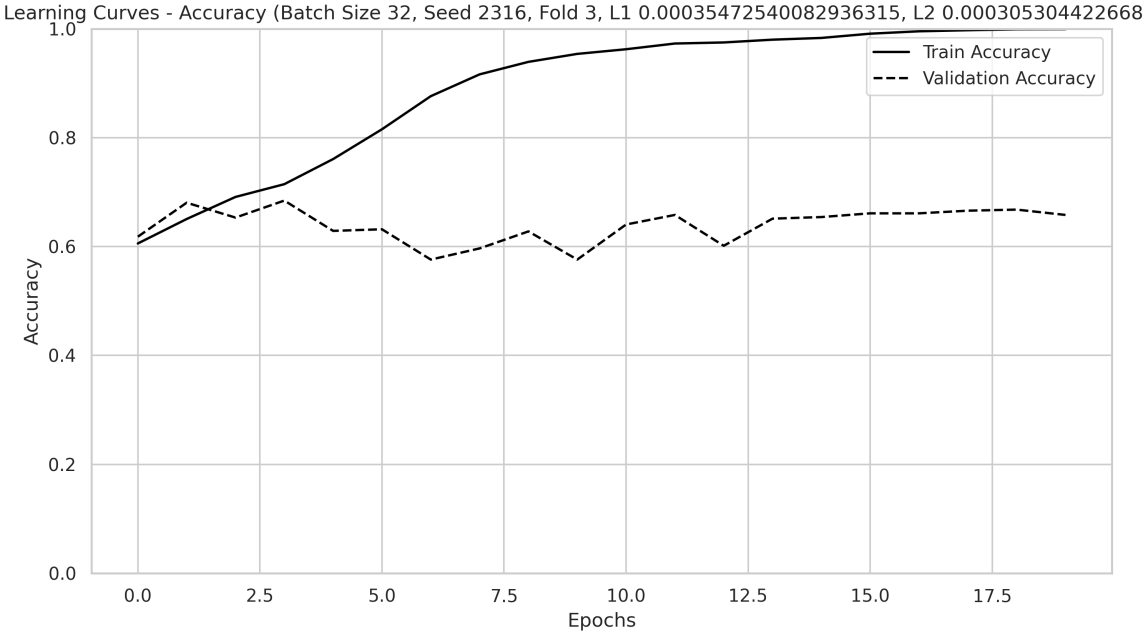


Figure 15: Demonstrating Over-fitting for the InceptionV3 and ResNet18 Hybrid Model

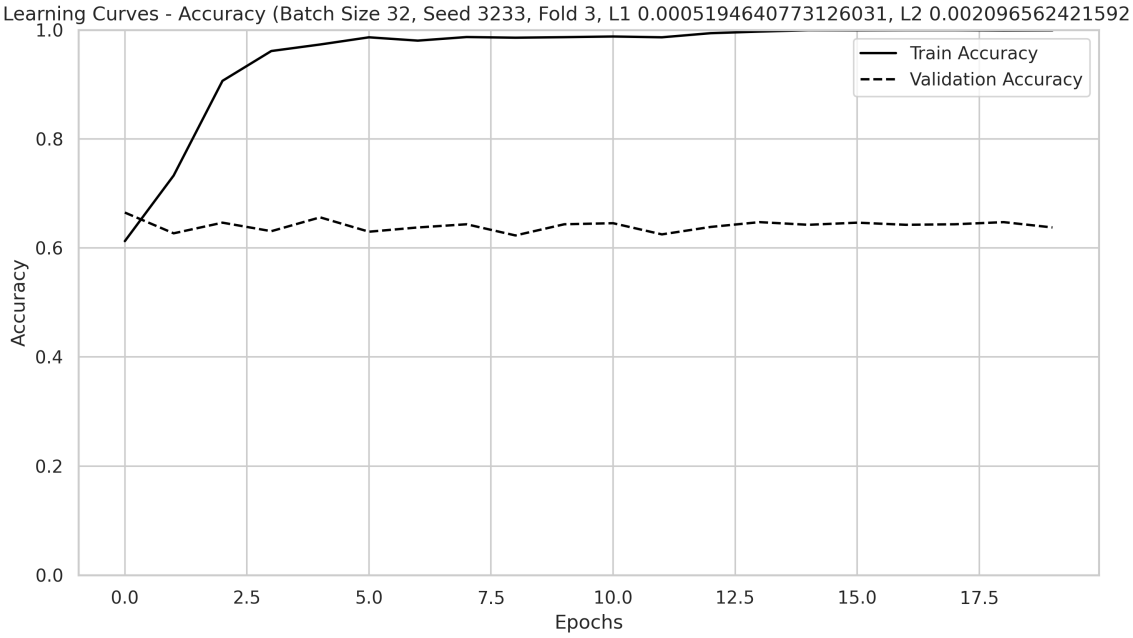


Figure 16: Demonstrating Over-fitting for the ShuffleNetV2 and ResNet18 Hybrid Model