

Application of Deep Convolutional Neural Network in Breast Cancer Prediction Using Digital Mammograms

by

Rafsan Al Mamun

18301033

Gazi Abu Rafin

21241072

Adnan Alam

21241071

MD. Al Imran Sefat

21241076

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science/Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
January 2022

© 2022. Brac University
All rights reserved.

Declaration


It is hereby declared that

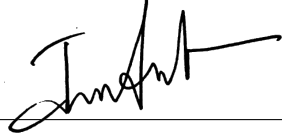
1. The thesis submitted is our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:


Rafsan Al Mamun
18301033


Gazi Abu Rafin
21241072


Adnan Alam
21241071


MD. Al Imran Sefat
21241076

Approval

The thesis titled “Application of Deep Convolutional Neural Network in Breast Cancer Prediction Using Digital Mammograms” submitted by

1. Rafsan Al Mamun (18301033)
2. Gazi Abu Rafin (21241072)
3. Adnan Alam (21241071)
4. MD. Al Imran Sefat (21241076)

Of Fall 2021 has been accepted as satisfactory in partial fulfilment of the requirement for the degree of B.Sc. in Computer Science and B.Sc. in Computer Science and Engineering on January 18, 2022.

Examining Committee:

Supervisor:
(Member)



Faisal Bin Ashraf
Lecturer

Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)



Moin Mostakim
Lecturer

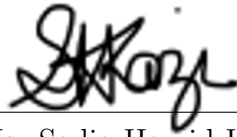
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam
Associate Professor

Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)



Ms. Sadia Hamid Kazi
Chairperson
Department of Computer Science and Engineering
Brac University

Abstract

Cancer, a diagnosis so dreaded and scary, that its fear alone can strike even the strongest of souls. The disease is often thought of as untreatable and unbearably painful, with usually, no cure available. Among all the cancers, breast cancer is the second most deadliest, especially among women. What decides the patients' fate is the early diagnosis of the cancer, facilitating subsequent clinical management. Mammography plays a vital role in the screening of breast cancers as it can detect any breast masses or calcifications early. However, the extremely dense breast tissues pose difficulty in the detection of cancer mass, thus, encouraging the use of machine learning (ML) techniques and artificial neural networks (ANN) to assist radiologists in faster cancer diagnosis. This paper explores the MIAS database, containing 332 digital mammograms from women, which were augmented and preprocessed, and fed into a custom and different pre-trained convolutional neural network (CNN) models, with the aim of differentiating healthy tissues from cancerous ones with high accuracy. Although the pre-trained CNN models produced splendid results, the custom CNN model came out on top, achieving test accuracy, AUC, precision, recall and F1 scores of 0.9362, 0.9407, 0.9200, 0.8025 and 0.8572 respectively while having minimal to no overfitting. The paper, along with proposing a new custom CNN model for better breast cancer classification using raw mammograms, focuses on the significance of computer-aided detection (CAD) models overall in the early diagnosis of breast cancer. While a diagnosis of breast cancer may still leave patients dreaded, we believe our research can be a symbol of hope for all.

Keywords: breast cancer, malignant, benign, mammogram, CAD model, convolutional neural network, convolution layer, overfitting, MIAS database, accuracy, precision, recall, F1, ROC curve, AUC

Dedication

We dedicate this research project to all the women who have lost their lives to breast cancer, to all those mothers who have glued together families selflessly without much caring about their health, to all those young ladies who are shut down from raising concerns about sensitive issues, to all those girls who feel hesitant to talk about their physical and mental troubles with others, to all those males who think only their female counterparts should be worried about breast cancer and not them, and finally, to every person in the world competing in a constant rat race to emerge at the top, not realising that it is their health that should be of utmost priority to them.

Acknowledgement

First and foremost, we would like to show our gratitude to Allah the Almighty, who has given us guidance and His blessings for which we could finish our research work on time with utmost diligence.

Secondly, we would like to extend our deep appreciation to our supervisor, Mr. Faisal Bin Ashraf and co-supervisor, Mr. Moin Mostakim who have shown us the way throughout the research and helped us finalise it.

Last, but not the least, we thank our parents who have supported us and believed in us for the whole of our lives, and without whom we would have never even reached the stage we are in right now.

Table of Contents

Declaration	i
Approval	ii
Abstract	iv
Dedication	v
Acknowledgement	vi
Table of Contents	vii
List of Figures	ix
List of Tables	x
Nomenclature	xii
1 Introduction	1
1.1 What Is Cancer?	1
1.1.1 Breast Cancer and its Severity	1
1.1.2 Problems of Diagnosing Breast Cancer	2
1.1.3 ML and its Use in Breast Cancer Diagnosis	3
1.2 Problem Statement	3
1.3 Research Objectives	5
1.4 Paper Orientation	5
2 Background	7
2.1 Convolutional Neural Network	7
2.2 Building Blocks of CNN Architecture	8
2.2.1 Convolution Layer	8
2.2.2 Non-linear Activation Function	9
2.2.3 Pooling Layer and Max Pooling	9
2.2.4 Fully Connected Layers	10
2.2.5 Last Layer Activation Function	10
2.3 Training a Network	10
2.3.1 Loss Function	11
2.3.2 Gradient Descent	11
2.3.3 Adam Optimiser	11
2.4 Data and Ground Truth Labels	12

2.5	Overfitting	12
2.6	Transfer Learning	13
2.7	Why Use CNN for This Research?	14
3	Literature Review	15
3.1	Research Using Clinical Data	15
3.2	Research Using Digital Mammograms	18
3.3	Major Findings and Scope of Research	23
4	The Dataset	24
4.1	Data Collection	24
4.2	Data Analysis	24
4.3	Data Preprocessing for Use in Models	26
4.3.1	Data Segmentation	26
4.3.2	Data Augmentation	27
5	The Models	29
5.1	Transfer Learning Models	29
5.1.1	ResNet50	30
5.1.2	MobileNetV3-Small	31
5.1.3	VGG19	32
5.2	Custom CNN Model	33
5.3	Model Evaluation	34
5.3.1	Confusion Matrix	34
5.3.2	Accuracy	35
5.3.3	Loss	35
5.3.4	Training vs Validation Accuracy and Loss	35
5.3.5	Sensitivity and Specificity	35
5.3.6	Precision, Recall and F_1 Score	36
5.3.7	ROC Curve and AUC	36
6	Result Analysis	38
7	Discussion	43
8	Conclusion	46
	Bibliography	51

List of Figures

1.1	Dividing Breast Cancer Cell [31]	2
1.2	Region-Specific Incidence and Mortality Rates for Female Breast Cancer, 2020 [57]	2
1.3	General Overview of the Proposed Solution	5
2.1	Overview of the Basic Structure of a CNN [44]	7
2.2	Convolutional Operation With Zero Padding [44]	8
2.3	Common Activation Functions Used in CNN [44]	9
2.4	Operation of a Max Pooling Layer [44]	10
2.5	Division of Dataset for Training a Model [44]	12
2.6	Fixed Feature Extraction of Transfer Learning [44]	14
3.1	Comparison of AUC Between ANN and Radiologists as Calculated in [15]	17
3.2	Comparison of AUC for Unaided Vs Aided Mammogram Analysis [51]	18
3.3	Mammogram Sample With Labelled Data From [35]	19
3.4	ROC Curve of the Model Used in [42] With AUC	20
3.5	Confusion Matrix Analysis for (a) ResNet50 and (b) VGG16 [52]	22
3.6	ROC Curves of the Models Trained on (a) DDSM and (b) INbreast [52]	22
4.1	Sample Mammograms from MIAS Dataset [4]	26
4.2	Sample Mammogram Before and After Augmentation	28
5.1	Dataset Segmentation for Training and Testing Models	29
5.2	Architecture of ResNet50	31
5.3	Architecture of MobileNetV3-Small [50]	32
5.4	Architecture of VGG19 [47]	33
5.5	Architecture of Custom CNN Model	34
5.6	Sample ROC Curve	36
6.1	ROC-AUC of the Models	39
6.2	Training Vs Validation Metrics of the Models	40

List of Tables

2.1	Common Last Layer Activation Functions for Particular Tasks 56	10
3.1	Accuracy of Different Models used in [43]	15
3.2	Accuracy Comparison of Methods from Papers Cited in [28]	16
3.3	Accuracy of Different Models Used in [39]	17
3.4	Comparison of Results Between the Models Used in [35]	20
4.1	Variables Used and Their Instances From the MIAS Database [4]	25
4.2	Distribution of the Study Population in MIAS Dataset [4]	25
6.1	Test Accuracy, Loss and Training Times of the Models	38
6.2	Average Precision, Recall and F_1 Scores of the Models	42

Nomenclature

Adam	Adaptive Moment Estimation
AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under Curve
BI-RAD	Breast Imaging - Reporting and Data System
CAD	Computer Aided Detection
CC	Craniocaudal
CLAHE	Contrast Limited Adaptive Histogram Equalisation
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CV	Cross Validation
DCGAN	Deep Convolutional Generative Adversarial Network
DDSM	Digital Database for Screening Mammography
DMR	Research Data Base
DNN	Deep Neural Network
FFDM	Full-field digital mammography
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GPU	Graphics Processing Unit
HIPAA	Health Insurance Portability and Accountability Act of 1996
IEEE	Institute of Electrical and Electronics Engineers
kNN	k-Nearest Neighbour
LDA	Linear Discriminant Analysis

LVQ	Learning Vector Quantisation
MIAS	Mammographic Image Analysis Society
ML	Machine Learning
MLO	Mediolateral Oblique
MQSA	Mammography Quality Standards Act
NB	Naïve Bayes
OpenCV	Open Source Computer Vision
PEIPA	Pilot European Image Processing Archive
ReLU	Rectified Linear Unit
ResNet	Residual Neural Network
RF	Random Forest
RGB	Red Green Blue
RIAC	Rule Induction through Approximate Classification
RMSprop	Root Mean Square Propagation
ROC	Receiver Operating Characteristic
ROI	Region(s) of Interest
SGD	Stochastic Gradient Descent
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
TPR	True Positive Rate
VGG	Visual Geometry Group
WDBC	Wisconsin Diagnosis Breast Cancer
WHO	World Health Organisation

Chapter 1

Introduction

1.1 What Is Cancer?

Cancer is one of the leading causes of death among the world population keeping the life expectancy prisoned to its barriers throughout the globe [56]. According to the research approved by the World Health Organisation (WHO), in 2020 alone, it claimed an estimated 10 million lives [55], [58].

In human body, cell division is a natural process where normal cells divide to replace the dead cells or replicates for growth. However, when cancer develops, this orderly process breaks down, resulting in the cells to become abnormal: old or damaged cells survive when they should die, and new cells form when they are not needed. These extra cells can divide without stopping and may form growths called tumours, which are divided into two types: malignant and benign. Malignant are the ones that spread and invade the other cells and organs, otherwise known as cancerous, whereas benign ones grow in its place into a large mass [31].

1.1.1 Breast Cancer and its Severity

Breast cancer is one such type of cancer. It occurs when the cell tissues of the breast become abnormal and uncontrollably divides. These abnormal cells form large lump of tissues, consequently becoming a tumour [16], [36]. Figure 1.1 shows a dividing breast cancer cell [31].

Breast cancer is considered as one of the most frequent non-skin cancers that develops in women, such that it is estimated that 1 out of every 8 women will develop it in their lifetime, making it one of the main causes behind women death [12], [36]. According to WHO, in 2020, breast cancer had the highest number of new cases (2.26 million), making it the most common type of cancer, surpassing lung cancer. It also reported 685,000 deaths, with ratio of 1 in 3 breast cancer patients who succumb to death [58]. Figure 1.2 below depicts the region specific incidence and mortality age-standardised rates for female breast cancer in 2020 [57].

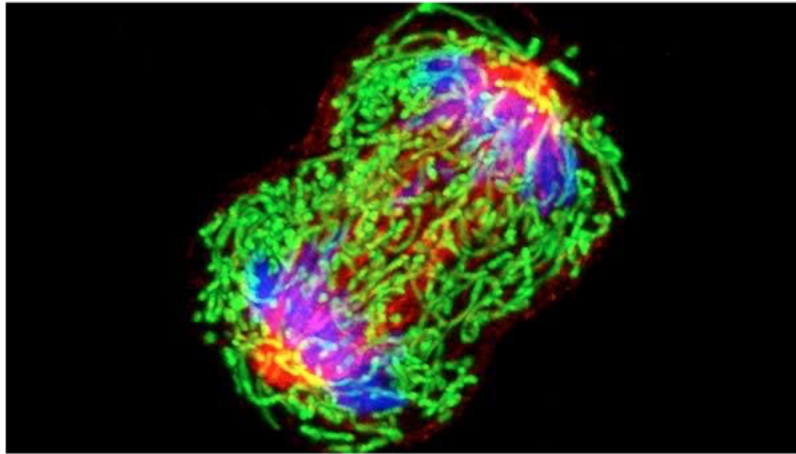


Figure 1.1: Dividing Breast Cancer Cell [31]

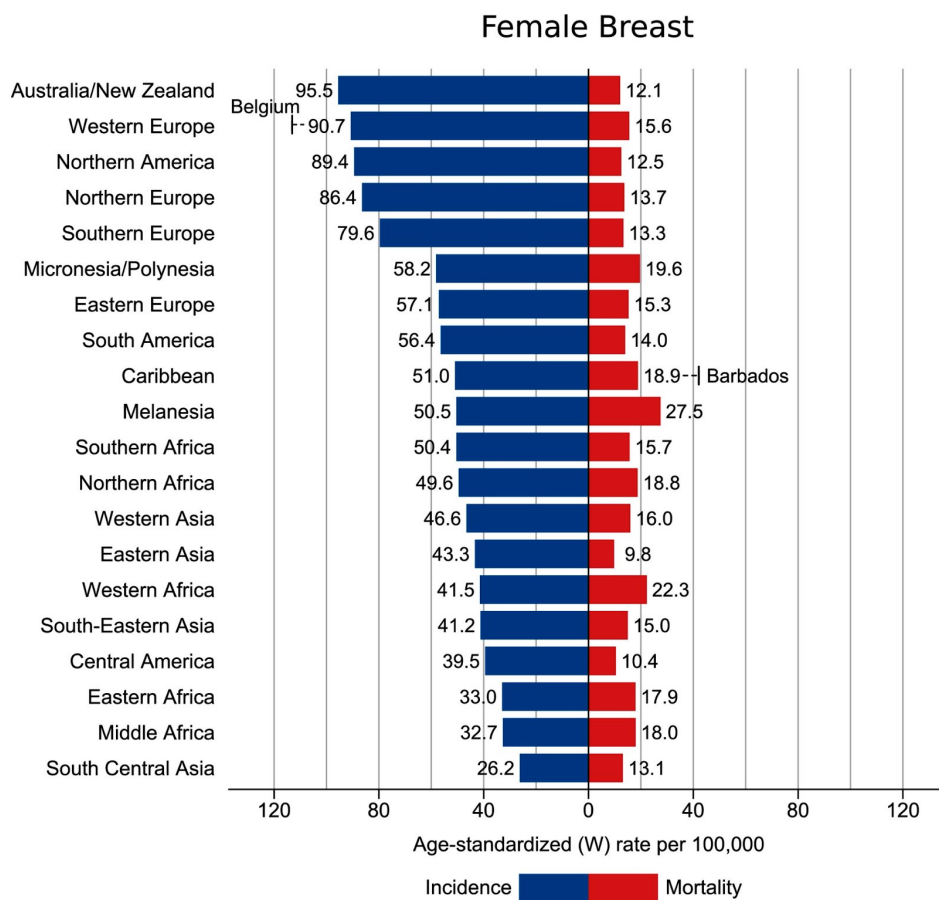


Figure 1.2: Region-Specific Incidence and Mortality Rates for Female Breast Cancer, 2020 [57]

1.1.2 Problems of Diagnosing Breast Cancer

Even though breast cancer is the most frequent malignancy in women worldwide, it is curable in $\sim 70\text{--}80\%$ of patients with early-stage, non-metastatic disease or phase, but is difficult to cure if it reaches distant organ metastasis [46]. However, as we have observed the ratio of deaths to new found cases, it boils down to the fact on how it was not diagnosed sooner when it was still in its early-stage. This again

reflects certain social aspects, such as in the developing countries where the education regarding this matter is neither that high nor is it encouraged, especially since it is a sensitive matter and most women are not comfortable to talk or think about it.

However, the biggest problem associated with misidentification of breast cancer is the inability to detect malignant breast tumours from mammography. The detection of masses by the mammography and making a classification and proper image out of it is very challenging particularly because of the varying size and density of female breasts. Radiologists, almost all the time, have issues with it because only 3% of the required information to determine the tumour and its type from normal tissue is being recovered from a mammogram [36]. This is a vital obstacle because, as stated earlier, breast cancer can be treated if it is detected early.

1.1.3 ML and its Use in Breast Cancer Diagnosis

Because of the difficulty of identifying breast cancer from mammograms by radiologists, it is necessary to develop and employ newer techniques to effectively and efficiently discover them. One such technique is using ML to make an artificial intelligence (AI) model that can detect such subtle features and apply its vast knowledge that it accumulated from training through a large data set to find the cancer cells effectively [36]. These can even detect the differences in various micro-calcification deposits or clusters of calcium in soft breast tissues- and masses which are very similar and are hard to distinguish, which usually form from the early onset of breast cancer [12].

A particular ML technique, known as CNN, uses a group of multiple neurones in various layers, taking on image type data, to find necessary patterns in them. It is designed to automatically and adaptively learn spatial hierarchies of features through back-propagation by using multiple building blocks, such as convolution layers, pooling layers, and fully connected layers [44]. These convolutional layers create feature maps that record a region of image which is ultimately sent out for nonlinear processing.

Therefore if a CNN model is trained with these raw mammograms, it should be able to detect the various patterns in the abnormality of a cancerous tissue compared to that of the normal tissue with higher accuracy in regards to other current techniques. Hopefully, these will give doctors a better chance to diagnose and treat patients with breast cancer from earlier onset.

1.2 Problem Statement

In this paper, the research focus is to classify a patient as cancerous (breast cancer) or healthy, $Res(t)$ which takes on the values 1 (patient has breast cancer) and 0 (the patient is healthy) based on, the probability that a patient has breast cancer, $P(C)$, given a digital mammogram, M of the patient and threshold, t , where t is the cutoff value of the probability beyond which the mammogram will be classified as cancerous. This is more formally defined by:

$$Res(t) = \begin{cases} 1; & P(C|M \cap t) \geq t \\ 0; & P(C|M \cap t) < t \end{cases}; \quad 0.00 \leq t \leq 1.00 \quad (1.1)$$

Cancer is the second major cause of death globally, accounting for about 1 in 6 deaths [57]. Among all the types of cancer, breast cancer is the most prevalent with about 7.8 million alive women with breast cancer diagnosis at the end of 2020. Women with breast cancer loses the most disability-adjusted life years than with any other types [58]. The risk of developing breast cancer by women is so critical, especially towards the later life, that even if all the potential risk factors, including avoidance of alcohol and tobacco, avoidance of using hormones, weight control etc, it would only reduce the risk of the onset of breast cancer by a mere 30% [58]. Hence, according to medical practitioners around the world, early detection is the most effective measure of minimising the risks from breast cancer, with mammography being the best cancer screening test at the moment [45].

However, interpreting mammograms pose great difficulties for the radiologists, as most breast cancers can go undetected. This is mainly because women have breasts of varying densities, and the fact that breast tumours are too small to be detected in the early stages. Typically women with dense breasts are more likely to have false-negative results, giving them a false sense of security [48]. Thus, there is a rising need for CAD models to help radiologists have a better diagnosis.

The literature review revealed the mass use of different ML models to differentiate between cancerous and healthy breasts with good accuracy, by using clinical data extracted from the mammograms by radiologists. Albeit other papers have utilised digital mammograms in detecting malignant tumours with the help of CNN, they used pre-identified regions-of-interest (ROI) [33], including lesions, to only target a part of the images. However, in order to build a better CAD model to try help the radiologists identify malignant tumours in breast mammography, it is adamant that the model makes use of the raw digital images of the the mammogram during the identification process, instead of specific ROI and being overly-dependent on radiologists. A model with such capabilities would help detect many cancerous tissues, that could have otherwise gone unnoticed by radiologists.

In this paper, the main focus is to accurately identify breast cancer using CNN by utilising raw mammograms, without any prior intervention by radiologists or other medical personnels. Figure 1.3 gives an overview of the solution that is being proposed in this paper. The aim is to show the capability of AI in breast cancer detection, after being trained with a large enough dataset, that would have otherwise gone unnoticed and postulate the significance of better CAD models to help radiologists with early detection of breast cancer, possibly saving thousands of lives in the process.

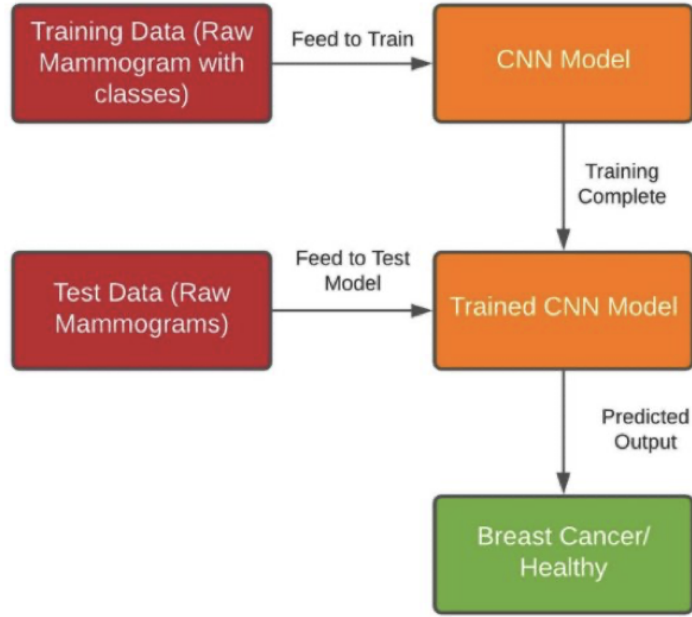


Figure 1.3: General Overview of the Proposed Solution

1.3 Research Objectives

In order to build a CNN capable of accurately detecting breast cancer among patients, it is necessary to set a standard procedure for attaining it. This would not only allow to assess the model based on a particular standard, but would also remove biases from the evaluation of the performance of the model.

In this research paper, digital mammograms from the MIAS database [4] have been extracted and augmented, to be fed into different CNN models in order to identify malignant breast tissues without the use of any pre-identified ROI as inputs. Apart from using already-existing pre-trained CNN models, another CNN model that was tailor-made for solving this particular problem would also be used to compare the results. For the purpose of discrimination, the models would be evaluated based on their accuracy of predicting healthy and cancerous breast tissues, precision, recall, F_1 score and the area-under-curve (AUC) of their receiver operating characteristic (ROC) curve plotted using their respective 1-specificity and sensitivity at different threshold values.

The CNN model that has the best above-mentioned metrics, along with lower loss and minimal overfitting, would be proposed for use in detecting breast cancer. The intention is to assert the significance of CNN-based CAD model using raw digital mammograms in the early and better diagnosis of breast cancer.

1.4 Paper Orientation

This chapter mainly introduces the readers to breast cancer, and provides a brief discussion of the research’s problem statement and objectives. The remainder of the

paper is organised as follows: **Chapter 2** provides an overview into the background knowledge for this paper, introduces the concept of CNN and justifies the use of CNN for this research. **Chapter 3** contains the literature review of some past published researches on the use of ML and deep learning in classifying healthy and cancerous tissues in breast cancer patients. **Chapter 4** focuses on the dataset, its analysis and preprocessing for being able to be used in the research. **Chapter 5** introduces the models that were used in this research, with their results being analysed in **Chapter 6**. **Chapter 7** provides a brief discussion about the probable factors that helped the custom CNN achieve supremacy, some limitations of the research and their proposed improvements. **Chapter 8** summarises the whole research and concludes the paper. Finally, there is a **Bibliography** at the end that lists out all the sites and journals that were referred to in this paper.

Chapter 2

Background

2.1 Convolutional Neural Network

Convolutional Neural Network (CNN) is a neural network model which falls under deep learning which processes data by using a grid pattern, i.e. images, developed by analysing the organisation of the visual cortex of animals [1], [2] which helps it to find differences in low to high level feature patterns.

A CNN has three components: convolution layer, pooling layer and fully connected layers. The task of the convolution and pooling layer is to extract features and can be run multiple times to extract additional features. The extracted features are passed on to the fully connected layers so that it can be mapped on to the output for functions like classification. The more layers the CNN has the more progressively complex the outputs become.

The model is trained for optimisation using different optimisation algorithms like backpropagation, gradient descent etc so that the outputs produced are more consistent with the “ground truth” labels. Figure 2.1 gives an overview of the basic structure of a CNN [44].

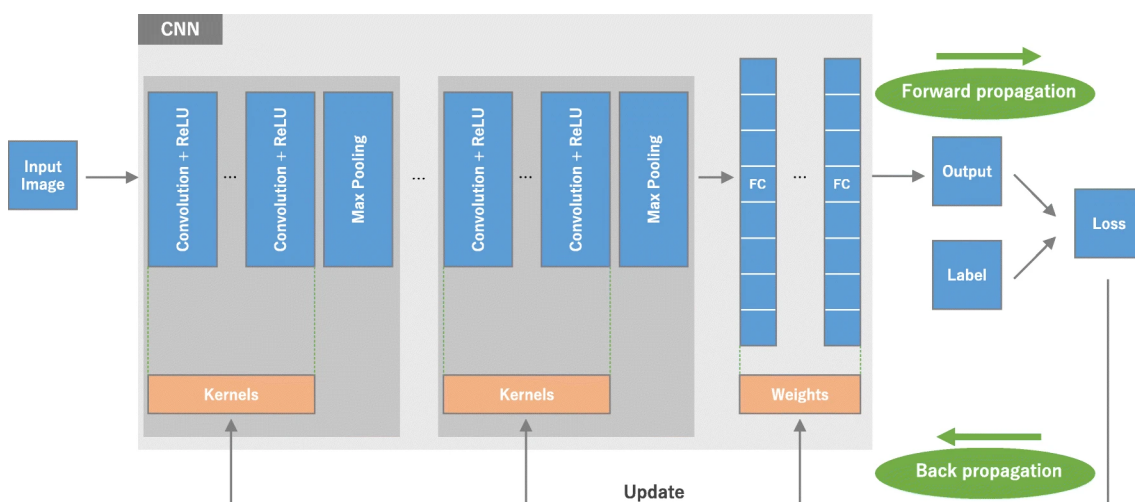


Figure 2.1: Overview of the Basic Structure of a CNN [44]

2.2 Building Blocks of CNN Architecture

A typical CNN is made up of several convolutional and pooling layers, connected to fully connected layers. A typical architecture usually consists of repetitions of several convolution layers and pooling layers, followed by one or more fully connected layers. The step where input data are transformed into output is called forward propagation.

2.2.1 Convolution Layer

The convolution layer plays a vital role in extraction of features by using linear and non-linear mathematical operations such as convolution operation and activation function.

Convolution is the mathematical combination of two functions to produce a third function. It merges two sets of information. In CNN, it is used by applying a small array of numbers known as a kernel or filter on the input data known as tensors. The elements of the kernel and tensor take part in an element-wise product specific to the individual location of the tensor and summed which results in the creation of a feature map which consists of outputs having unique positions in the output tensor. Moreover, multiple different kernels are applied to the input tensor to extract different features from the datasets.

The convolution operation has two main arguments and they are size and number of kernels. Most common kernel size is 3×3 while 5×5 and 7×7 are also seen. Conventionally, convolutional operations do not allow the centre of the kernel to overlap the outermost element of the input tensor meaning the loss of data in the feature map, which can be countered by employing a technique called zero padding. Figure 2.2 shows a convolution operation with zero padding [44].

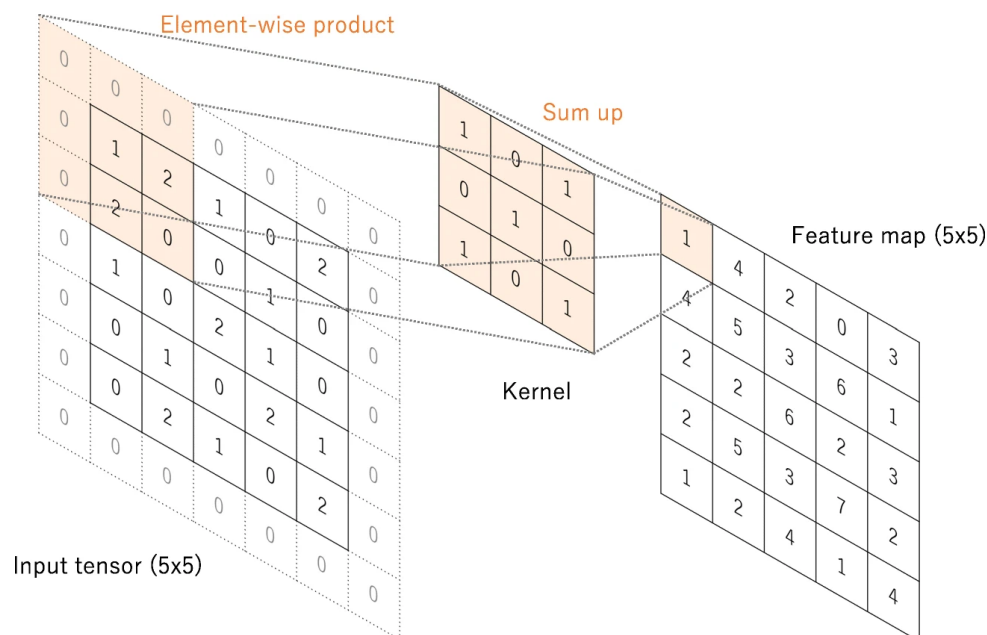


Figure 2.2: Convolutional Operation With Zero Padding [44]

A stride is the distance between two successive kernel positions and is commonly 1. However, a larger one is sometimes used to downsample the feature maps.

Weight sharing refers to the kernels being shared between image positions is an important factor of the convolutional operation. This allows for the local features detected by one kernel to be used as variables by other kernels so no time is wasted detecting local features again. Spatial hierarchies of feature patterns can be learned by downsampling and using a pooling operation, which results in a larger feature map. Moreover, the efficiency of the model can be increased by reducing the number of parameters that are required.

2.2.2 Non-linear Activation Function

The feature map produced from the convolution layer is later passed on to a non-linear activation function. The most common activation function is the rectified linear unit (ReLU). Figure 2.3 below shows some common activation functions used in the inner layers of a CNN [44].

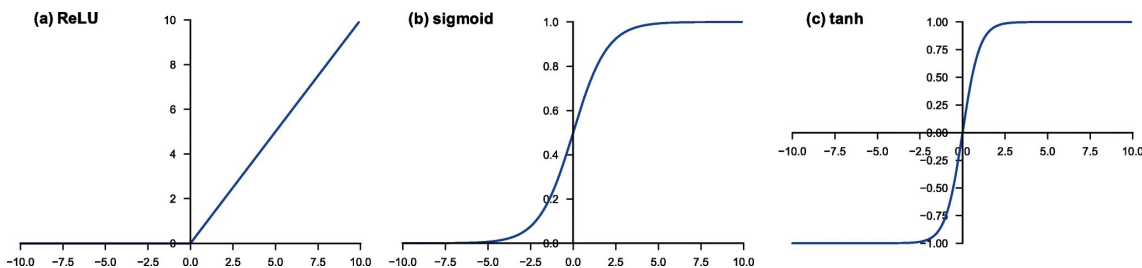


Figure 2.3: Common Activation Functions Used in CNN [44]

2.2.3 Pooling Layer and Max Pooling

The pooling layer is used for downsampling the feature map's size in order to introduce a translation invariance which detects small shifts and distortions, and helps to decrease the number of parameters learned during training.

Max pooling is the favoured pooling operation where the feature map is split into groups of patches and the highest value is chosen from each patch and the rest get discarded as shown in Figure 2.4 [44]. The filter size 2×2 with a stride of 2 is the most popular max pooling. This downsamples the feature map by a factor of 2.

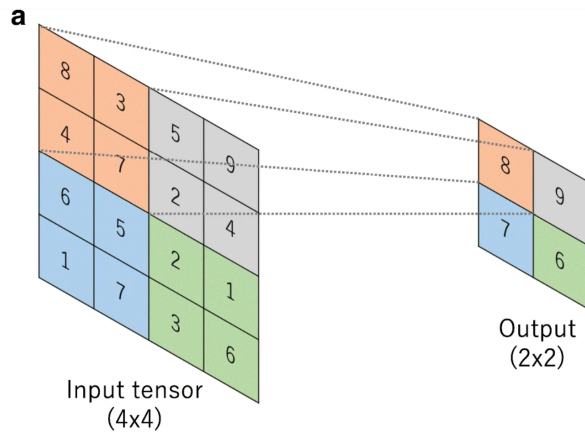


Figure 2.4: Operation of a Max Pooling Layer [44]

2.2.4 Fully Connected Layers

The outputs of the convolutional and pooling layers are connected to one or more fully connected layers in which every input is connected to every output by a learnable weight. Once the features have been extracted by the convolution layer and been downsampled by the pooling layer, they are passed on to fully connected layers which produce the final outputs like probabilities for classification etc. The number of classes determine the number of output nodes in the final fully connected layer. Each layer is followed by a nonlinear function, such as ReLU, as described above.

2.2.5 Last Layer Activation Function

The activation function used in the last layer is different to the functions used in the other layers and the function used is dependent on the task. Table 2.1 below shows the last layer activation function that are typically used for some particular tasks [44].

Task	Last Layer Activation Function
Binary Classification	Sigmoid
Multiclass Single-class Classification	Softmax
Multiclass Multi-class Classification	Sigmoid
Regression to Continuous Values	Identity

Table 2.1: Common Last Layer Activation Functions for Particular Tasks **56**

2.3 Training a Network

A network is trained to find a combination of unique kernels in the convolution layer and weights in the fully connected layers that produce outputs with minimum difference from the labeled dataset used. Backpropagation is the main algorithm used for training neural networks with hidden layers which uses the loss function and gradient descent optimisation algorithm, among others.

2.3.1 Loss Function

Loss function is a function that calculates the difference between the actual output and the output from the network by using forward propagation, and this is labeled as the cost. The most popular loss function for multiclass classification is cross-entropy and for regression to continuous value, mean squared error is used. The one used in this paper is the binary cross-entropy.

2.3.2 Gradient Descent

Gradient descent is an optimisation algorithm whose main function is to minimise the loss by regularly updating the learnable parameters like kernels and weights in the network. However, the algorithm that will be used in this paper is known as Adaptive Moment Estimation (Adam), which is more of an improvement upon the general gradient descent algorithm.

2.3.3 Adam Optimiser

Adam is an adaptive learning rate optimisation algorithm and is often referred to as a combination of two optimisation algorithms which are Root Mean Square Propagation (RMSprop) and Stochastic Gradient Descent (SGD) with momentum [23]. This is said because, to scale the learning rate, it squares the gradients like RMSprop and uses the moving average of the gradient instead of the gradient like SGD with momentum. Being an adaptive learning rate method means different parameters result in different learning rates. Adam uses adaptive moment estimation, and so to adapt the learning rates of each weight, it uses the estimates of the first and second moments of the gradient. Adam also keeps an exponentially decaying average of past gradients. These are done by using the following formulae [23]:

$$\hat{m}(t) = \frac{m(t)}{1 - \beta_1(t)} \quad (2.1)$$

$$\hat{v}(t) = \frac{v(t)}{1 - \beta_2(t)} \quad (2.2)$$

where, $m(t)$ = first moment, $v(t)$ = second moment, β_1 and β_2 = hyperparameter, t = batch number.

The typical values for β_1 and β_2 are 0.9 and 0.999 respectively. In order to update the weight, Adam uses the following formula:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}(t) + \varepsilon}} \hat{m}(t) \quad (2.3)$$

where, θ = weight, η = learning rate, ε = zero-avoidance parameter = $10e^{-8}$.

2.4 Data and Ground Truth Labels

In any machine learning methods or deep learning, datasets and ground truth labels are the most important content. In fact the success of any such method and models are dependent on its dataset and ground truth label. Therefore, it is most necessary to carefully select the datasets and ground truth labels, but then again, obtaining such high quality ones is both expensive and time consuming [44]. As for medical images, there are multiple good quality sources readily available. However, to be used for a specific topic or specific function, the model needs data sets with particular ground truth labels and hence, special care needs to be taken.

Usually datasets are of 3 categories: a training, validation and test set. As the name suggests, the training set is used to train the network, where loss values are calculated via forward propagation and learnable parameters are updated back into the network via backpropagation. Validation set is used for fine-tuning the hyperparameters and performing model selection during the whole training process. At the very end, the final model or network is run through the test set and its final performance after all those tuning using training and evaluation datasets is evaluated (see Figure 2.5 [44]). It is notable that evaluation and test sets are kept different. This is particular because the training model's hyperparameters are fine-tuned according to the performance it showed while using the evaluation set.

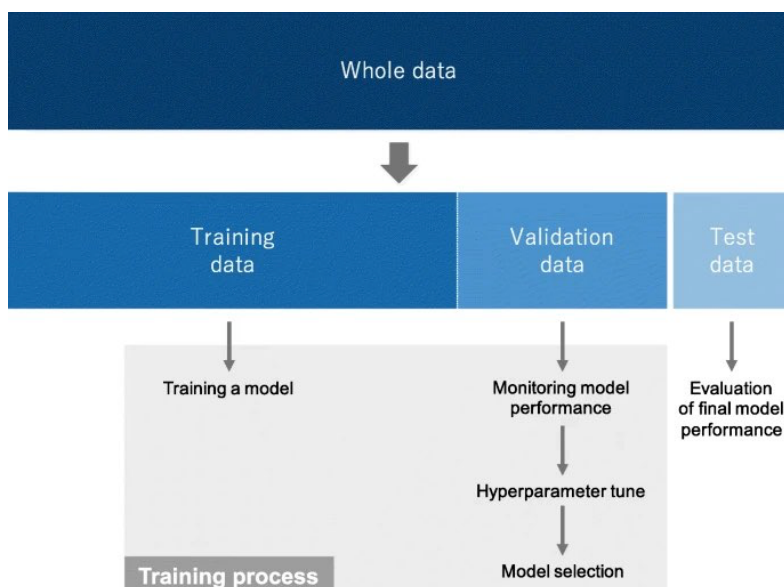


Figure 2.5: Division of Dataset for Training a Model [44]

2.5 Overfitting

Overfitting is the phenomenon when a model learns the statistical regularities specific to the training set, in other words it learns the unnecessary information or noise particular to the dataset instead of the signal, hence performing poorly on the new dataset. Overfitting has always been a challenge and thus, the test set is

used to evaluate the performance of the model. Regular checkup to recognise the overfitting to the training data by monitoring the loss and accuracy to the training and validation sets is usually done [44].

Although there are solutions to avoid these in practice, the best solution to avoid overfitting of course is to have more training data. However, given such is not always available, there are other ways such as regularisation with dropout, weight decay, data augmentation etc [44].

Dropout is a regularisation technique, where randomly selected activations are set to 0 during the training, so that the model becomes less sensitive to specific weights in the network [18]. Weight decay or L2 regularisation penalises the model's weights so that it takes only small values. Batch normalisation is a type of supplemental layer, which standardises the input values of the following layer for each mini batch adaptively, thus, reducing overfitting. It has the effect of stabilising the learning process and dramatically reducing the number of training epochs required to train a deep network. Data augmentation, on the other hand, is a process where the training data is modified through random transformations, such as flipping, translation, cropping, rotating, random erasing etc, so that the model sees different input while training [37], [44].

2.6 Transfer Learning

Although large datasets are highly desired in training a model, such datasets are hard to find. One way to mitigate this problem is to use transfer learning, as it trains the network model on a large dataset, like ImageNet, then reuses the pre-trained model for the topic of interest. The assumption that is made is features learned on a large dataset can be shared among seemingly disparate datasets [44]. This ability to shift the learned generic features from datasets is what gives deep learning the advantages to make itself useful in various domain tasks with small datasets. Some examples of such models are AlexNet, VGG, ResNet etc.

While there are many ways to use the pre-trained network, this paper will focus on fixed feature extraction. A fixed feature extraction method is a process to remove fully connected layers from a network pre-trained on a large database, while maintaining the remaining network, which consists of a series of convolution and pooling layers, referred to as the convolutional base, as a fixed feature extractor (see Figure 2.6 [44]). The fixed feature extractor can further be topped off with fully connected layers in CNN resulting in training limited to the added classifier on a given dataset of interest [44].

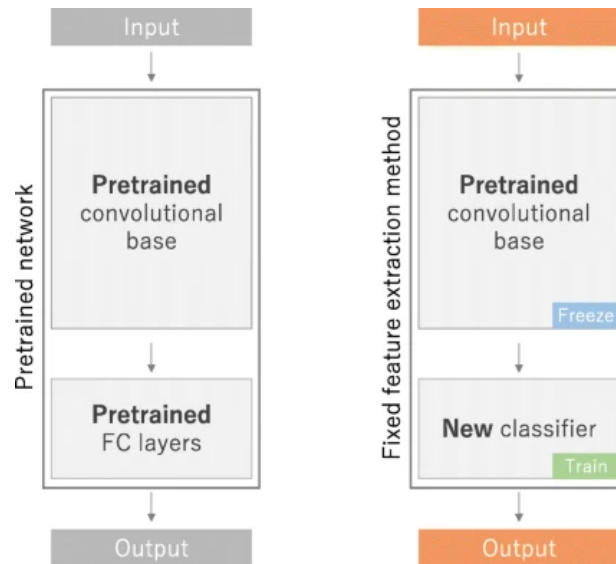


Figure 2.6: Fixed Feature Extraction of Transfer Learning [44]

2.7 Why Use CNN for This Research?

Over the years, CNNs have developed much rapidly, thanks to its ability to accurately conduct difficult classification functions in images that would have otherwise required abstract concepts. What gives CNN an edge compared to its old competitors is its ability to detect important features in a dataset without the need for any human supervision, making it the go to model for a lot of industrial systems. As stated before, the core concept of CNN is that it uses different convolutions and filters to produce invariant features that are passed on to the next layer, where more new filters and convolutions are applied to extract further features, until it gives a final output.

However, the key feature of CNN is that it works well on image data, as the several convolution layers derive benefit from the fact that an interesting pattern can occur anywhere in an image, in contiguous blocks of pixels, allowing it to learn useful features from raw data without manual image processing. Since the purpose of this research is to predict breast cancer from raw image data with complex features, CNN is more than capable of standing up to the task, producing better predictions than any other models. Hence, CNN was chosen to be used for this research.

Chapter 3

Literature Review

As stated earlier, to be able to better facilitate subsequent medical treatments, it is really necessary to be able to detect malignant breast tissues. Breast cancer is usually detected after the conductance of a special type of X-ray, called a mammography, which is later scrutinised by radiologists for a better diagnosis. However, it has been revealed that in about 1 in 5 breast cancers get undetected by radiologists after screening [48]. Hence, it is understood that CAD techniques, especially ML, could aid radiologists to have a better interpretation of the patients. Thus, a number of past papers that have tried to incorporate such techniques have been reviewed to find the scope of further research and have been summarised below.

3.1 Research Using Clinical Data

Most papers have used clinical data gathered from mammograms as a means of detecting breast cancer. One such paper [43] used the commonly used ML techniques, namely Random Forest (RF), k-Nearest Neighbour (kNN) and Naïve Bayes (NB). For reference, the author has cited different other papers: Detection using Relevance Vector Machine [32] yielded results with an accuracy of 97%, using the Wisconsin Diagnosis Breast Cancer (WDBC) dataset [3]; Mamdani Fuzzy inference model for training was used in one study in conjunction with Linear Discriminant Analysis for feature selection and acquired an accuracy of 93% [26]. The paper [43] utilised the WDBC dataset [3], consisting of 569 instances attributed mainly by diagnosis, mean radius, mean texture, and mean area of the ROI, split it into 10 different chunks using the 10-fold cross validation (CV) method, and fed them into said ML models. The accuracy of the models can be found in Table 3.1 [43]. Although the author produced great accuracy exceeding 94% in breast cancer identification, the dataset was too small to be used to train non-parametric models like NB, increasing chance of overfitting.

Method	RF	kNN	NB
Accuracy (%)	94.74	95.90	94.47

Table 3.1: Accuracy of Different Models used in [43]

Similar to the one above, another paper [28] has also utilised the WDBC dataset [3], albeit an updated one with 699 instances. The study proposed the use of a new NB technique, called a weighted NB, to classify breast tissues as malignant or benign. The dataset used had 9 different clinical features, namely clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell, bare nuclei, bland chromatin, normal nucleoli and mitosis, for each of the patients. The author used a similar technique as [43] to validate the model, known as the 5-fold CV. However, since the author used a weighted NB, using weights to control the attributes in a posterior probability contribution meant that changes should be done to the original NB equations to properly represent them, bringing a big disadvantage of assigning crisp classes to the training data [13]. The author further used a heuristic search algorithm to calculate the weights to be used. After training and testing the model, the author achieved an average accuracy of 98.54%. Table 3.2 depicts a comparison of the author’s model with other papers the author has cited [28]. Even though the author was able to manipulate the NB classifier to achieve greater result than others, while doing so, the paper introduced a heuristic search algorithm that is computationally expensive on which the model is dependent to obtain the weights, along with a bias that cannot provide generalised result when used on other datasets.

Study	Method	Accuracy (%)
Hamilton et al. [5]	RIAC	94.99
Ster and Dobnikar [6]	LDA	96.80
Bennett and Blue [7]	SVM	97.20
Setiono [9]	Neuro-rule	98.10
Goodman et al. [10]	Big-LVQ	96.80
This study [28]	W-NB	98.54

Table 3.2: Accuracy Comparison of Methods from Papers Cited in [28]

It seems like the WDBC dataset [3] is a popular choice among researchers trying to discriminate breast cancer patients from healthy ones using ML techniques. Another recent paper [39] utilised the same dataset for use in NB and kNN classifiers for breast cancer classification. The authors of this paper split the dataset into training and testing set in the ratio of 60:40 respectively, before feeding it to a standard kNN model with k=3. For the NB, the authors calculated the mean and standard deviation of each feature for each category (malignant and benign), which were then used to calculate the probability for each prediction the model made. However, the model had a poorer accuracy than kNN as seen in Table 3.3 [39]. This research, apart from suffering from small dataset, might have also suffered from under-training since only 60% of the dataset has been used solely for training, thus reaching a lower accuracy than [28] despite resorting to a similar model, albeit a simpler one.

Method	kNN	NB
Accuracy (%)	97.51	96.19

Table 3.3: Accuracy of Different Models Used in [39]

The use of clinical features extracted from mammograms has not only been used in typical ML models, but also in ANNs, in order to achieve even better results. A paper from the year 2010 [15], was one such paper that made use of a feed-forward ANN, with 36 discrete input variables, split using the 10-fold CV method, and a hidden layer with 1000 nodes. The paper not focused on discriminating the benign cells from the malignant cells, but also on stratifying patients into high and low risk groups. The paper worked with over 60000 mammography findings matched with the Wisconsin State Cancer Reporting System, consisting of a 5-level BI-RADS assessment [11]. The ANN, trained with early stopping method to avoid overfitting, achieved a significantly higher mean AUC (0.965) compared with that of radiologists (0.939; $p < 0.001$) calculated from their respective ROC curves as seen in Figure 3.1 [15]. The authors evaluated the accuracy of their model's risk prediction by using the Hosmer-Lemeshow goodness-of-fit test [21], which showed a high calibration. The results demonstrated that this model may have the potential to help the radiologists in discriminating between benign and malignant breast tissues. The difference of the AUC between the ANN and the radiologist may look small (0.026) but this difference is significant both statistically and clinically as the model identified 44 more cancers and decreased the number of false positive by 3941 when compared with the radiologists. However, the one flaw in this research was the fact that the authors depicted a BI-RAD score of 0 as positive, whereas in reality, the score 0 means unsure, hence, causing some inaccuracies in their result.

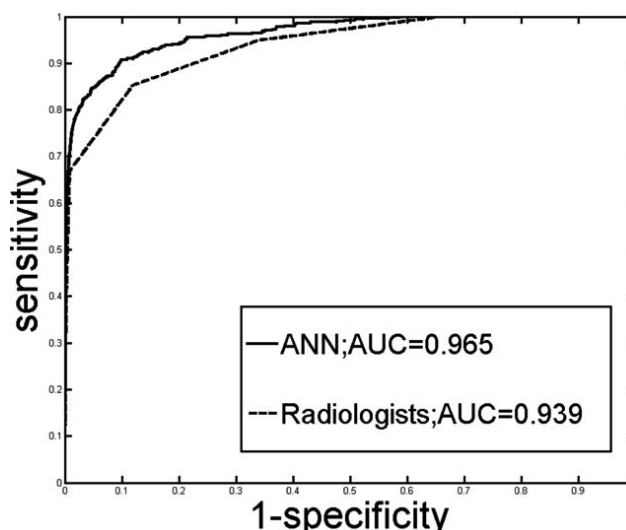


Figure 3.1: Comparison of AUC Between ANN and Radiologists as Calculated in [15]

3.2 Research Using Digital Mammograms

The papers cited in the previous section mainly centred their attempts of building CAD models to help in the accurate detection of breast cancer among patients using primarily the clinical data extracted from the mammograms and the patients. However, another way of detecting breast cancer is by analysing the mammographic images using a CNN without the need of other clinical inputs. This is more likely to ensure a better CAD model that do not rely on radiologists' interpretation, but rather try to find the specific patterns in the images by itself.

One such paper [51] that incorporated the idea of CNN for breast cancer classification, compared the breast cancer detection performance of the radiologists reading the mammograms unaided versus supported AI systems. This was an enriched retrospective, fully crossed, multi-reader, multi-case, HIPAA-compliant study, which used digital mammographic examinations from 240 women. The mammograms were later interpreted by 14 Mammography Quality Standards Act (MQSA) - qualified radiologists, with and without AI support. The system used deep CNN, features classifiers and image analysis algorithms to depict calcifications and soft-tissue lesion in two separate modules, which were then combined to determine suspicious region findings. These regions were later given values between 1 and 100 which represented the level of suspicion that breast cancer was present (with 100 indicating the highest suspicion). Figure 3.2 shows the difference between the AUC of the ROC curves for the two reading conditions, compared by using mixed-models analysis of variance and generalised linear models for multiple repeated measurements [51]. The paper could have been an excellent case for showing the power of AI in the detection of breast cancer and its superiority over radiologists. Unfortunately, however, the results were not much promising as the AUC difference between the radiologists' investigation with and without AI support was a mere 0.02, with $p=0.002$ (the difference is not significant). Hence, it calls for better CAD models that can analyse raw images to classify breast cancer patients.

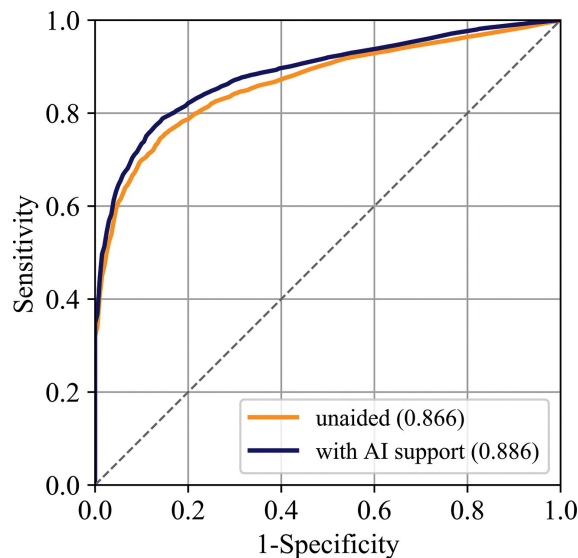


Figure 3.2: Comparison of AUC for Unaided Vs Aided Mammogram Analysis [51]

Another paper that has studied the detection of breast cancer from mammograms by utilising a CNN but with much better results than [51] was the paper from [35]. However, before diving into building the models arbitrarily, the author has into previous works in order to grasp the different types of breast cancer and how they are classified with the use of multiple kinds of ANN that specialise in image classification. Firstly, the CNN tutorial on TensorFlow was used in order to test the functionalities of the features it offers [24]. Secondly, for image classification, a model from [19] which was the ImageNet Classification with deep CNN was looked at and the preprocessing techniques present were used as references in [35]’s work. The dataset used in the paper was from the mini-MIAS, containing 322 grayscale mammograms with labelled data, describing the type of cancer, properties of the background tissue, class of abnormality and the coordinates of the centre of abnormality. The authors further used image transformations techniques in order to augment the dataset. A random mammogram sample from this dataset has been shown in Figure 3.3 [35]. Unlike other researches, the authors used three different versions of CNN to assess their results: one was the CovNet model from Kaggle [34], second CNN model was developed following the TensorFlow tutorial [24] that takes the whole image as input with labelled data, and the third used a 48x48 input matrix, convolution layer with kernel size 5x5 filter with ReLU activation, pooling layer with max pool size 2x2 filter and a stride value of 2, learning rate with 0.003, and training step with 20,000 samples. Their results have been tabulated in Table 3.4 [35], which shows the third version surpassing the other two with 82.71% accuracy. Although the model could have been improved by removing the labelled data from the images, it still shows the capabilities of AI to help the doctors in this field to correctly identify breast cancer, if present, in the patient faster.

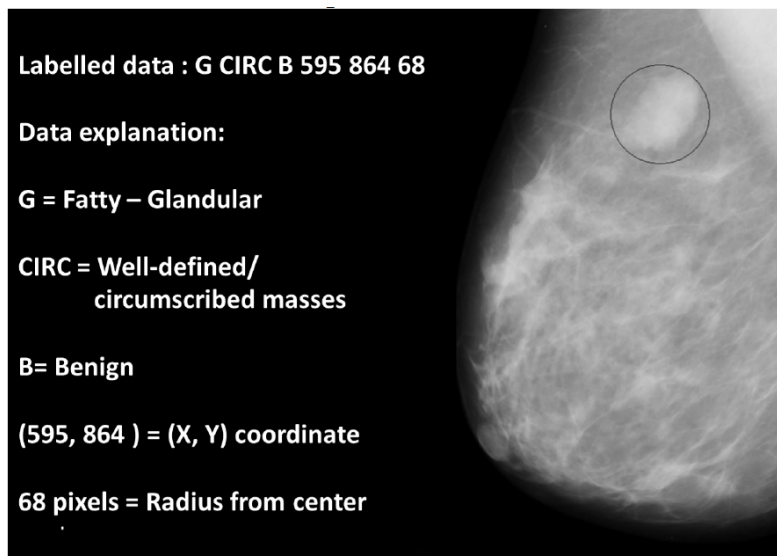


Figure 3.3: Mammogram Sample With Labelled Data From [35]

Version	Version 1	Version 2	Version 3
Sensitivity (%)	35.32	54.32	82.68
Specificity (%)	35.43	54.36	82.73
Accuracy (%)	38.45	54.35	82.71

Table 3.4: Comparison of Results Between the Models Used in [35]

When searching for more studies which used neural networks for classification of mammograms, two further studies of interest were found that met the search criteria. The first one [20] used an ANN and the second one [22] used a wavelet neural network to identify breast cancer. However, both of these methods have a lot of parameters (~ 3 million) when compared to a normal CNN (~ 600) as they were made for making general decisions.

Next, there was one paper [42] that was of particular interest because of its use of an unorthodox methods of detecting breast cancer from images. This study first explored infrared imaging, that assumes that there is always an increase in thermal activity in the precancerous tissues and the areas surrounding developing breast cancer. The study used the Research Data Base (DMR) database containing frontal thermogram images, including breasts of various shapes, from 67 women. The thermograms went through image pre-processing to mark the ROIs and remove the unwanted regions like arms, neck etc. They were then fed into an AI model built using a deep neural network (DNN) together with a Support Vector Machine (SVM) model as classifier. The SVM is only consulted with if the DNN is incapable of classifying the images with great confidence. The model which is presented in this paper takes advantage of two main factors. First, a DNN (pre-trained Inception V3 model [30], [40]) which is modified at the last fully connected layer in such a way as to obtain a powerful binary classification which can tell if a cell is healthy or cancer infected. Secondly, a well known classifier (SVM) is coupled to that and is involved only in the case of uncertainty in the output of the DNN. The results of this study is presented in Figure 3.4 [42] that shows an AUC of 1.00 calculated from its ROC curve. Even though the study has shown a perfect accuracy of detecting breast cancer, such models should be dealt with extreme caution as, ironically, such accurate models could also be a result of overfitting, which could be particularly true in this case because of the extremely small dataset used in the study.

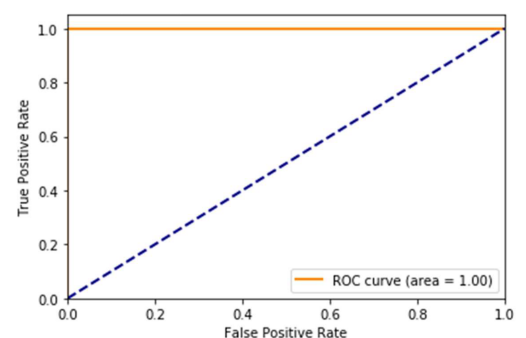


Figure 3.4: ROC Curve of the Model Used in [42] With AUC

Finally, the best paper in this category, in terms of the sophistication of the model used and the accuracy attained, [52] had a goal of detecting breast cancer from screening mammograms by using a deep learning algorithm, trained by an “end-to-end” approach, allowing the datasets to be either complete clinical annotations or only the labeled cancer region in the image. Many studies before it applied deep learning models, but these models were used to classify annotated lesions because ROI in a mammogram is too small when compared to a full-field digital mammography (FFDM). Some studies were also found that used unannotated FFDM datasets to train neural networks, but the results were inconclusive [38], [41].

The dataset used in this study [52] were taken from the DDSM database which included a total of 2478 mammography images taken from 1249 women. The mammograms consisted of standard views such as craniocaudal (CC) and mediolateral oblique (MLO) and these views were used as separate images. The images consisted of annotations for the ROIs, the type of cancer and whether it was a mass or calcification. The sampling of image patches from the ROIs and background regions resulted in a lot of images which were later split into two datasets: S1 and S10. S1 consisted of a mix of image patches which focused on the centre of ROIs and random background patches from each of the images, while S10 had 10 patches randomly selected from the regions surrounding the ROI coupled with background patches to paint a big picture of the ROI. These patches were further classified as background, malignant mass, benign mass, malignant calcification and benign calcification. This dataset was used in the pre-training of the classifier. Another dataset from the INbreast database was also used in this study. This dataset had 410 mammograms with CC and MLO views from 115 patients. The mammograms had radiologists’ BI-RADS assessment which were: 1-no findings, 2-benign, 3-probably benign, 4-suspicious, 5-highly suggestive of malignancy and 6-biopsy-proven cancer. The images with BI-RADS 3 were excluded and BI-RADS 1 and 2 were labeled as negative and 4, 5 and 6 as positive. This dataset was used to train the whole image classifier.

This study made the model in such a way that it needed a fully annotated lesion dataset only during pre-training for the initialisation of the weight parameters of the model, and then labeled images without ROI annotations could be used for the rest of the training. This is beneficial as large databases of annotated lesions are hard to come by. For pre-training usually a two step method is used which is made up of classification of the annotated ROIs by a classifier which generates a grid of probabilistic outputs and these outputs are summarised to find out the classification of the outputs to their respective classes. The author suggested a new method which combines both of these steps in order to optimise it. This is done by using the input patches found in the images and putting them directly into a CNN classifier instead of a conventional classifier. The output from the CNN is a grid of probabilistic outputs of the classes instead of it being classified into single classes.

Two CNN structures were used in this study: the VGG network [29] and the residual network (ResNet) [27], specifically a 16-layer VGG network (VGG16) and a 50-layer ResNet (ResNet50) respectively. The results of using the VGG16 and ResNet50 on the DDSM dataset are shown as a confusion matrix in Figure 3.5 [52], that tell both

networks performed well overall but struggled with correctly identifying malignant classifications followed by correctly identifying malignant mass.

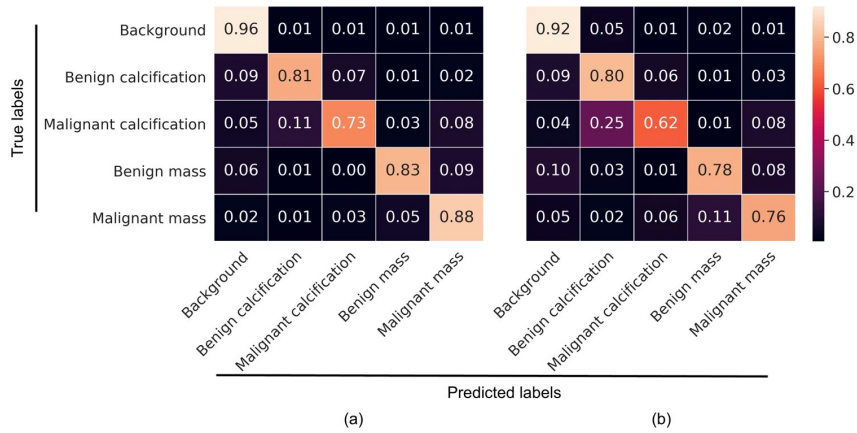


Figure 3.5: Confusion Matrix Analysis for (a) ResNet50 and (b) VGG16 [52]

Lastly, two hybrid networks have been created by first adding the best performing VGG blocks as top layer (two VGG block of $(256) \times 1$ and $(128) \times 1$) and the ResNet50 as bottom layers and vice versa (two blocks of $(512-512-1024) \times 2$). Moreover, to further efficiently train the networks, augmentation prediction was used which meant that the model trained on each image was flipped vertically and horizontally resulting in four images and the average AUC of the four images was used. After all this training, four models were identified as top performing and they were (patch classifier-top layer): ResNet-ResNet, VGG-VGG, ResNet-VGG and VGG-ResNet. This marked the end of the training of the model using annotated lesions. Next, the best models were trained on the INbreast dataset. The annotations on this dataset were ignored as the authors wanted to focus on the performance of the whole image classifiers on unannotated lesions and the transferability. The ROC graphs of these networks trained on the DDSM and INbreast datasets can be seen from Figure 3.6 [52]. The INbreast dataset produced a mean AUC of 0.95 (higher than DDSM) from all the four models. Moreover, using an ensemble of all the models resulted in an AUC of 0.98 and scored 86.7% and 96.1% in sensitivity and specificity respectively.

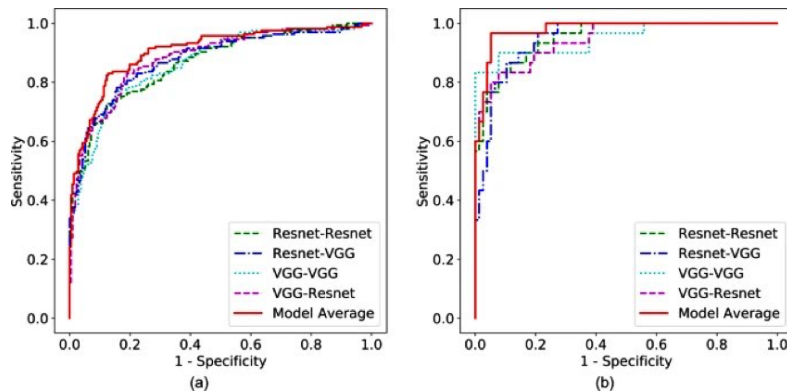


Figure 3.6: ROC Curves of the Models Trained on (a) DDSM and (b) INbreast [52]

However, trying to achieve such high accuracy comes with its prices. To begin with, it is found that to efficiently train the network during pre-training to provide optimum results, the network requires a large number of patches or large patches in general. Using large patches means that the computational cost also linearly increases. Furthermore, large patches require more GPU memory in order to process them properly. In addition, the whole image classifier was trained on the INbreast dataset which was labeled based on the BI-RADS provided by the radiologist, which has a chance of being wrong. The classifier being trained on it means it will incorporate a bias instead of finding new unknown characteristics in the images. Even so, nonetheless, this study shows the power of using end-to-end training deep learning models to produce highly accurate results in depicting breast cancer from mammograms, which can then easily be transferred to other datasets with little effort.

3.3 Major Findings and Scope of Research

After an exhaustive search through a corpus of past papers that had made use of ML techniques and deep learning for the classification of breast cancer, there were three main findings that would set the base for this paper:

1. Most researchers usually opt for clinical data extracted from mammograms as opposed to using the whole image to train the ML models for breast cancer identification. These assessments, especially the BI-RADS scores are subject to human-error, leading to inaccuracies in the results of the models.
2. Although some papers have tried to incorporate the idea of using images, most had used specific ROI as targets for the models to analyse, instead of trying to explore the whole mammogram for favourable features. This again causes the models to be over-reliant on radiologists' observation and speculation of a specific region that needs to be targeted.
3. Even if some papers have been able to achieve high accuracy and AUC from their models for breast cancer detection, most of them had used a relatively small dataset than that is required to effectively train the models. As a result, these models might not be proficient enough to generalise over unseen data.

The study in this paper is to be carried away in such a way as to minimise the effects of the above mentioned observations. The study intends to incorporate unannotated digital mammograms taken from the MIAS database [4] classified into normal, benign and malignant breast samples, which would be augmented to increase the size of the dataset for a better CAD model. CNN models are intended to be used to extract features from the dataset, in order to find patterns to aid in the detection of breast cancer patients.

Although a paper [35] tried to use raw images without specific ROIs in CNN, it failed to achieve better accuracy mainly due to the model's simplicity. The paper, that utilised the MIAS dataset, would be used as the comparison base for this research, with the intention to build a CNN model sophisticated enough to be able to handle the problem at hand, and ultimately help doctors detect breast cancers in patients early and with better confidence.

Chapter 4

The Dataset

4.1 Data Collection

All the mammograms from the database of the Mammographic Image Analysis Society (MIAS) [4], an organisation of UK research groups who are particularly interested in the understanding of mammograms, have been included in the dataset used in this paper for a retrospective evaluation. The films, compiled in the year 1994, were taken from the UK National Breast Screening Programme and had been digitised to 50 micron pixel edge with a Joyce-Loebl microdensitometer. The dataset was then further reduced to 200 micron pixel edge and had been clipped and/or padded to make all the images of size 1024x1024 pixels, hence the name, mini-MIAS. The final dataset is available via the Pilot European Image Processing Archive (PEIPA) at the University of Essex [4].

4.2 Data Analysis

The mini-MIAS database comprised of 328 digital raw mammograms belonging to 161 patients (image of left and right breasts of each individual). 6 of the mammograms were duplicates and so, have been removed to reduce the dataset to 322 samples. For each of the mammograms, the database contained the reference number, breast density, the abnormality present, severity of the abnormality, the x and y coordinates of the centre of abnormality, and the approximate radius of a circle enclosing the abnormality as can be seen in Table 4.1 [4].

When calcifications were present in the image, the coordinates and the radii apply to a cluster instead of individual calcification, with the bottom-left corner taken to be origin. Moreover, in some cases, the calcifications were distributed throughout the image rather than concentrating in a single site, for which the coordinates and radii are inaccurate and so, have been omitted.

The mammograms were prospectively analysed and interpreted by experienced radiologists who had between 1-35 years of experience in breast imaging. Their interpretations have been considered as the “ground-truth” values for the purpose of this research.

Variables	Instances
MIAS Reference Number	6-digit reference number
Breast Density	-Fatty (F) -Fatty-glandular (G) -Dense-glandular (D)
Abnormality Present	-Calcification (CALC) -Well-defined/circumscribed masses (CIRC) -Spiculated masses (SPIC) -Ill-defined masses (MISC) -Architectural distortion (ARCH) -Asymmetry (ASYM) -Normal (NORM)
Severity of Abnormality	-Benign (B) -Malignant (M)
Coordinates of Centre of Abnormality	x, y image-coordinates
Radius of Area of Abnormality	Radius (in pixels)

Table 4.1: Variables Used and Their Instances From the MIAS Database [4]

In order to properly analyse the dataset, the study population have been distributed based on certain criteria as can be found in Table 4.2 [4]. As can be seen, almost two-thirds (64%) of the total study population had normal breast tissues, with only 16% having malignant breast tumours and the rest having benign. Figure 4.1 shows random sample mammograms of patients with normal, benign and malignant breast tissues [4].

Study Population	Normal (%)	Benign (%)	Malignant (%)	Total (%)
No. of mammograms	207(64)	65(20)	50(16)	322(100)
Breast Density				
Fatty	66(20)	22(7)	18(6)	106(33)
Fatty-glandular	65(20)	22(7)	17(5)	104(32)
Dense-glandular	76(24)	21(6)	15(5)	112(35)
Abnormality Present				
Calcification	-	12(4)	13(4)	25(8)
Well-defined/ circumscribed masses	-	19(6)	4(1)	23(7)
Spiculated masses	-	11(3)	8(3)	19(6)
Ill-defined masses	-	7(2)	7(2)	14(4)
Architectural distortion	-	9(3)	10(3)	9(6)
Asymmetry	-	6(2)	9(3)	15(5)

Table 4.2: Distribution of the Study Population in MIAS Dataset [4]

After a careful analysis of the dataset, it can be seen that most of the study population (35%) had dense-glandular breasts. It is worth noting that, people with this density of breasts were least identified as having malignant breast tissues (5%), and most identified as having normal breast tissues (74%). Now, this could either

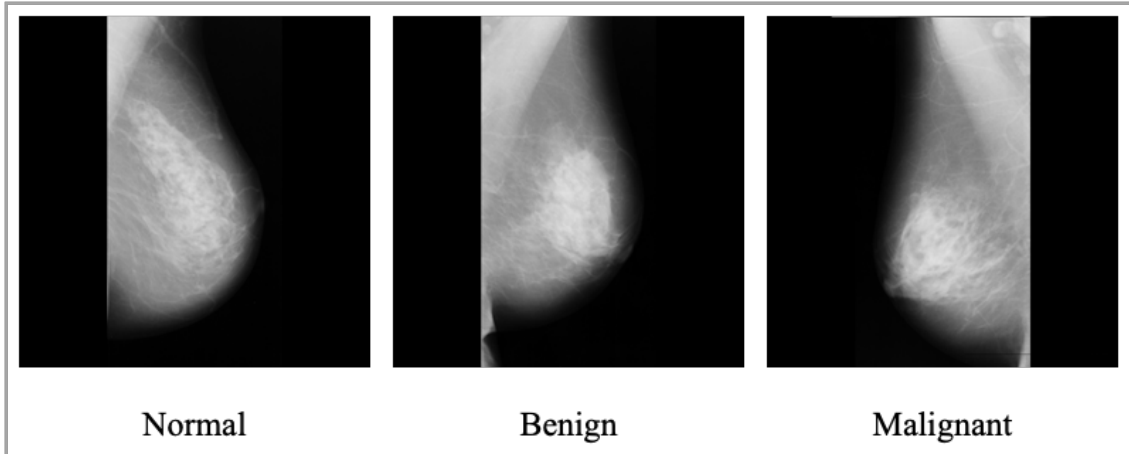


Figure 4.1: Sample Mammograms from MIAS Dataset [4]

be a normal phenomenon or it could re-iterate the fact that it is difficult to detect breast cancer in dense breast as stated earlier. Hence, there could be some level of discrepancies in the radiologists' assessment, but, since the difference is quite small compared to other breast densities, this has been neglected.

Moreover, it could also be seen that most breast tumours were in the form of calcification (8% among total and 22% among patients with tumour) and well-defined/circumscribed masses (7% among total and 20% among patients with tumour), while the least common form of tumour was ill-defined masses (4% among total and 12% among patients with tumour). It should also be noted that, among patients with a malignant tumour, almost 16% had a combination of fatty breast tissue and calcifications, which is the highest.

However, since the purpose of this research is to be able to predict breast cancer from raw digital mammograms, without relying on radiologists' clinical assessments and ROI, these clinical data have not been used in the models. Rather, only the raw image files and their interpreted labels have been incorporated in them.

4.3 Data Preprocessing for Use in Models

Before the images can be used in the CNN models, they need to be preprocessed in order to ensure that there are no bias or discrepancies in the models' predictions due to the nature of the data.

4.3.1 Data Segmentation

Since the research being done is to predict if a patient has breast cancer or not, the multiple segmentation of the dataset, i.e. normal, benign and malignant, are not needed to train the models. Hence, the dataset has been divided into 2 categories: "healthy" (patients having normal and benign breast tissue) and "cancer" (patients having malignant breast tissue). Therefore, the final dataset now had the original 322 images split into a ratio of 272:50 into the 2 categories respectively.

4.3.2 Data Augmentation

As stated early, one of the key improvements this research would have over others mentioned in the literature review, is to mitigate the disadvantage of working with a small dataset. When a neural network model is given a small-sized dataset to train, that model becomes overfitted, and memorises the data instead of the relationships. The goal of a model is to generalise patterns from the training data so that it can predict new data which was not available during training. The fewer the samples for training, the more models tend to become an overfitting model. Thus, the images need to be processed and augmented for the total dataset to increase in size. The preprocessing steps were done using the OpenCV [8] and Albumentations [54] library and are as follows (see Figure 4.2):

1. **Resize to 224x224 pixels:** The images were scaled down proportionally from the original 1024x1024 pixels to 224x224 pixels. This was mainly done to reduce the complexity of working with a large array of pixels, and also because the pre-trained models to be used in this research has an input size of 224x224 pixels.
2. **Shift from RGB channel to LAB colour space:** In order to apply Contrast Limited Adaptive Histogram Equalisation (CLAHE), images need to be turned to grayscale from RGB. However, since the input channel of most popular pre-trained models is RGB, the images had to be shifted to the LAB colour space (L, A, B stands for luminescence, red/green coordinates and blue/yellow coordinates respectively), so that CLAHE could be applied to the luminescence channel.
3. **Equalised using CLAHE and shifted back to RGB channel:** Contrast Limited Adaptive Histogram Equalisation (CLAHE) is a variant of Adaptive Histogram Equalisation which takes care of over-amplification of the contrast, and improves its quality [53]. CLAHE was applied on the images using a clip limit of 5. Later, the images were converted back to the RGB channel to be fed into the models. The resulting images after applying CLAHE were used as the base to apply the later augmentations on.
4. **Rotation:** The images were rotated anticlockwise by 10 and 20 degrees, to make the changes look subtle.
5. **Flip:** The images were then flipped both on the vertical and the horizontal axis.
6. **Random Tone Curve:** This method randomly changes the relationship between bright and dark areas of an image by manipulating its tone curve. This too was applied to ensure colour distortion of images were not an issue when making predictions.
7. **Gaussian Noise:** Gaussian Noise is a statistical noise having a probability density function equal to normal distribution. This noise was added to introduce some graininess in the images.

8. **Blur:** The images were also blurred by using a random-sized kernel, although the blurriness was kept tenuous to ensure minimal distortion of the original images.

After the completion of all the augmentation steps, the dataset, which once had only 322 mammograms, now contained 2898, with a total of 2448 mammograms belonging to the healthy category and the rest 450 to the cancer category.

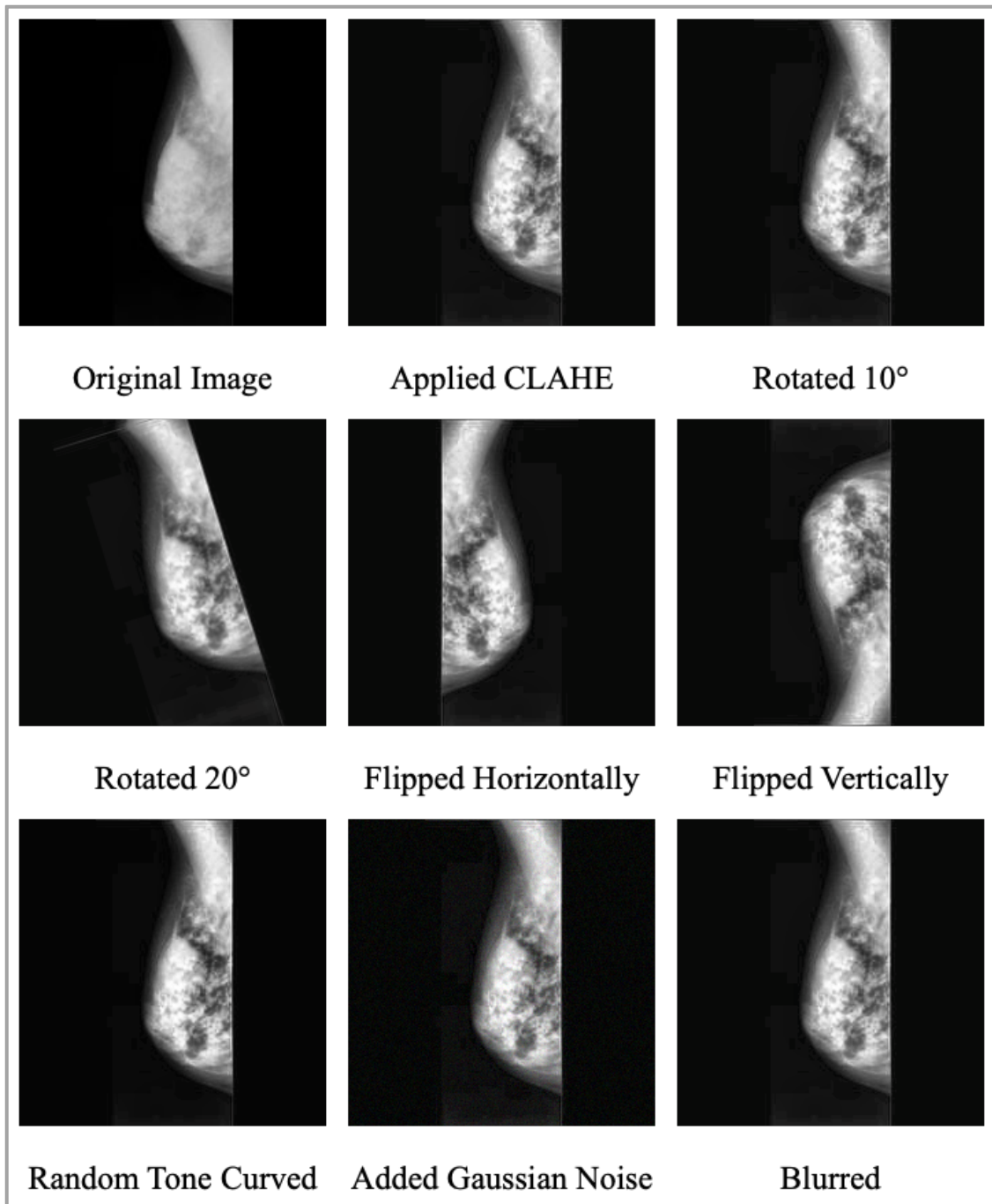


Figure 4.2: Sample Mammogram Before and After Augmentation

Chapter 5

The Models

After performing the preprocessing steps mentioned in Chapter 4, the dataset is ready to be fed into CNN models to establish a relationship between the features of the mammograms and the final interpretation. The dataset was further split in 80:20 ratio to be used for training and testing purposes respectively. 10% of the training data (8% of the total dataset) was used for validation while the rest for training the models. This was done using SciKit-Learn's [17] `train_test_split()` method using the parameter "shuffling=True" to ensure a mix of classes in each set of data used. This segmentation of the dataset would also allow to analyse if there is any overfitting of the models or not. Figure 5.1 shows the segmentation of the dataset used for training and testing the models.

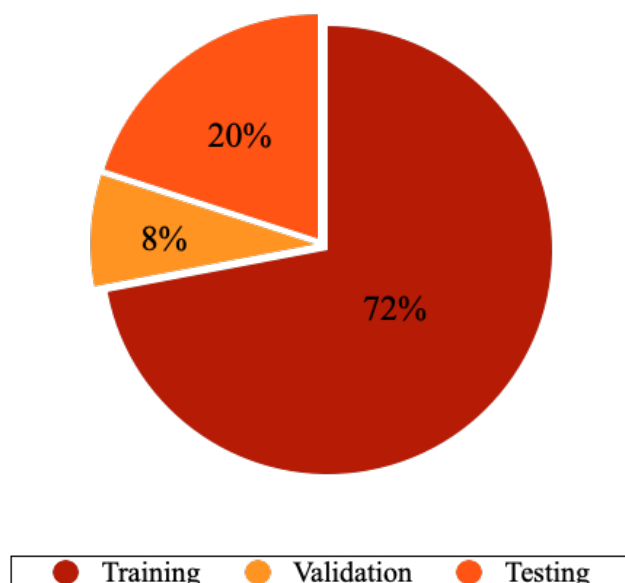


Figure 5.1: Dataset Segmentation for Training and Testing Models

5.1 Transfer Learning Models

As stated previously in Section 2.6, transfer learning allows the use of models pre-trained on a large dataset to be used for other classification problems, that make use of relatively smaller datasets. The transfer learning models used for this research

were trained on the ImageNet dataset [14], an image database that has been organised according to the WordNet hierarchy. Containing more than 14 million images, the database has been significant in advancing deep learning research. The models were implemented through the process of fixed feature extraction, by freezing all the layers to retain their learned weights, while replacing the last fully connected layer to fine tune them for this problem-specific tasks. Several models were tested upon using the TensorFlow library [25], and only the best performing ones were further exploited for better efficiency.

5.1.1 ResNet50

ResNet50 is a type of residual neural network (ResNet) model made up of 50 layers, out of which 48 layers are convolutional layers, 1 MaxPool and 1 AveragePool layer. ResNet is popular as it allowed the use of ultra deep neural networks which contained hundreds or thousands of individual layers with great performance [60]. Simply stacking layers on an existing network produces higher error, which ResNet overcomes by performing identity mappings using shortcut connections that skip one or two layers. Hence, one of the biggest advantage of it was that no additional parameters were added to the model while the computational time remained the same.

The ResNet50 architecture contains the following elements (see Figure 5.2):

1. A convolution with a kernel size of 7x7 and 64 kernels with a stride of 2.
2. A max pooling layer with a stride of 2.
3. Convolution layers with 64 1x1 kernels, 64 3x3 kernels and 256 1x1 kernels. These layers are then repeated 3 times.
4. Convolution layers with 128 1x1 kernels, 128 3x3 kernels and 512 1x1 kernels. These are then repeated 4 times.
5. Convolution layers with 256 1x1 kernels, 256 3x3 kernels and 1024 1x1 kernels, which too are repeated for 6 times.
6. Convolution layers with 512 1x1 kernels, 512 3x3 kernels and 2048 1x1 kernels, repeated 3 times.
7. An average pooling layer connected to a fully connected layer containing 1000 nodes, ending with a softmax activation function.

For the purpose of this research, the last fully connected layer was replaced by a fully connected layer (after flattening the previous outputs to a one-dimensional vector) with 512 hidden nodes, ReLU activation and a dropout of 0.5, before finally connecting to an output layer with 1 node and sigmoid activation function. Hence, the model now has more than 51 million trainable and more than 23 million non-trainable parameters. The model was compiled using Adam optimiser (learning-rate=0.001), binary-crossentropy as the loss function and accuracy as the metric. The model was trained and tested using various configurations, but the best average result was obtained by training the model for 8 epochs with 100 steps per epoch, repeating it for 5 times.

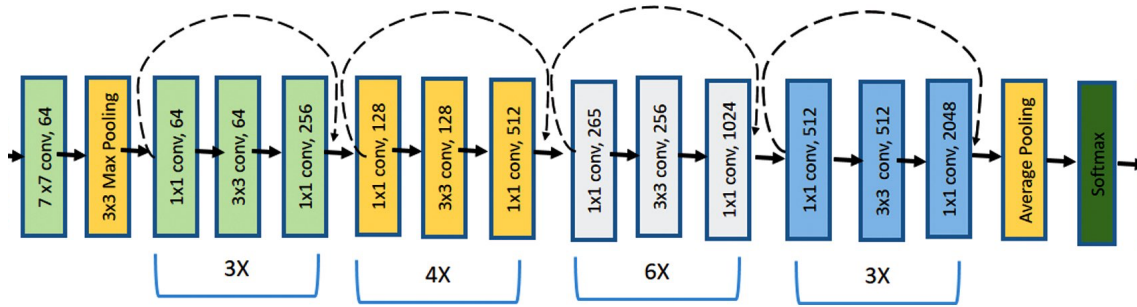


Figure 5.2: Architecture of ResNet50

5.1.2 MobileNetV3-Small

MobileNets are a family of CNNs, developed by Google, for the mobile phone and embedded architectures [49]. They are based on a streamlined architecture which makes use of depth-wise convolutions to build lightweight deep CNNs. MobileNetV3 has been tuned to the CPUs of mobile phones through a combination of network architecture search complemented by the NetAdapt algorithm. MobileNetV3-Small mainly targets low resource usage.

The layers in the MobileNetV3-Small architecture are as follows (see Figure 5.3 [50]):

1. A convolution layer with 16 3x3 kernels with stride of 2.
2. A bottleneck convolution layer with 16 3x3 kernels with stride of 2.
3. A bottleneck convolution layer with 24 3x3 kernels with stride of 2.
4. A bottleneck convolution layer with 24 3x3 kernels.
5. A bottleneck convolution layer with 40 5x5 kernels with stride of 2.
6. A bottleneck convolution layer with 40 5x5 kernels, repeated for 2 times.
7. A bottleneck convolution layer with 48 5x5 kernels, repeated for 2 times.
8. A bottleneck convolution layer with 96 5x5 kernels with stride of 2.
9. A bottleneck convolution layer with 96 5x5 kernels, repeated for 2 times.
10. A convolution layer with 576 1x1 kernels, followed by a pooling layer of 7x7 kernel.
11. A convolution layer with 1024 1x1 kernels, connected to a fully connected layer.

This time again, the last fully connected layer has been replaced by a fully connected layer (after flattening the previous outputs to a one-dimensional vector) with 512 hidden nodes, ReLU activation and a dropout of 0.5, before finally connecting to an output layer with 1 node and sigmoid activation function. The model has a total of more than 25 million trainable and 1.5 non-trainable parameters. After following the same technique for compilation as the previous model, the model was trained and tested using several constraints. The best average result came after training for 8 epochs with the default 66 steps per epoch, repeated for 5 times.

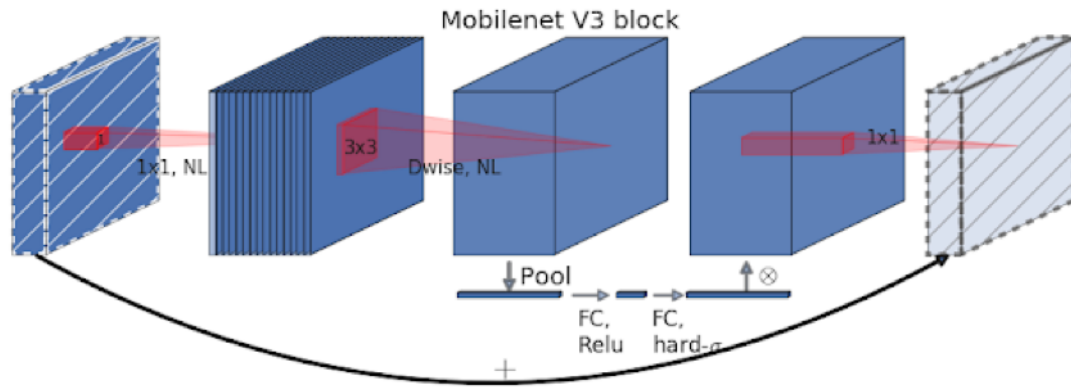


Figure 5.3: Architecture of MobileNetV3-Small [50]

5.1.3 VGG19

VGG19 is a variant of the Visual Geometry Group (VGG) model family, consisting of 16 convolution layers, 3 fully connected layers, 5 MaxPool layers and 1 SoftMax layer [59]. The model carries and uses some of the ideas from its predecessors and improves them, while also improving the accuracy using deep convolutional neural layers.

The layers in the VGG19 architecture are as follows (see Figure 5.4 [47]):

1. A convolution layer with 64 3x3 kernels, with a stride of 1, and repeated 2 times.
2. A convolution layer with 128 3x3 kernels, with a stride of 1, and repeated 2 times.
3. A convolution layer with 256 3x3 kernels, with a stride of 1, and repeated 4 times.
4. A convolution layer with 512 3x3 kernels, with a stride of 1, and repeated 4 times.
5. A convolution layer with 512 3x3 kernels, with a stride of 1, and repeated 4 times.
6. A MaxPool layer of 3x3 kernel size and stride of 2 after each group of convolution layers.
7. 2 fully connected layers with 4096 hidden nodes.
8. A fully connected layer with 1000 hidden nodes, followed by a softmax activation layer.

Similar to the previous model, the last fully connected layer of this model too has been replaced by a fully connected layer (after flattening the previous outputs to a one-dimensional vector) with 512 hidden nodes, ReLU activation and 0.5 dropout, before connecting to the output layer with sigmoid activation function. This allowed to build a model with almost 13 million trainable parameters and about 20 million



Figure 5.4: Architecture of VGG19 [47]

non-trainable ones. The compilation of the model used the same technique as all the models preceding it. After much trial-and-error, similar to MobileNetV3-Small, this model also yielded the maximum average results after training for 8 epochs with the default steps per epoch of 66, repeating it for 5 times.

5.2 Custom CNN Model

One of the main objectives of this research was to build a custom CNN, tailor-made to act as an effective CAD model in helping radiologists classify breast cancer among patients using raw mammograms without any human intervention or influences (such as ROI or clinical observations). While doing so, the plan was to aim for fast and simplistic models, with much less parameters to train but with higher accuracy than the already available pre-trained models used in this research.

The custom CNN that was built and trained had an architecture as follows (see Figure 5.5):

1. A convolution layer with 64 4x4 kernels, with a stride of 3, ReLU activation and an output shape of 74x74x64.
2. An average pooling layer of 3x3 kernel size, with a stride of 1 and an output shape of 24x24x64.
3. A convolution layer with 64 3x3 kernels, with a stride of 1, ReLU activation and an output shape of 22x22x64.
4. A max pooling layer of 2x2 kernel size, with a stride of 1 and an output shape of 11x11x64.
5. A convolution layer with 128 3x3 kernels, with a stride of 1, ReLU activation and an output shape of 9x9x128.
6. A max pooling layer of 2x2 kernel size, with a stride of 1 and an output shape of 4x4x64.
7. A fully connected hidden layer, with 256 hidden nodes, ReLU activation, 0.2 dropout and an output shape of 4x4x256.
8. A fully connected hidden layer, with 512 hidden nodes, ReLU activation, 0.2 dropout and an output shape of 4x4x512.

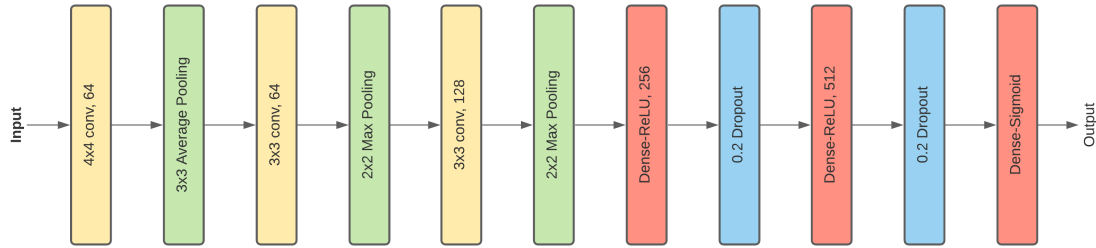


Figure 5.5: Architecture of Custom CNN Model

9. A final fully connected output layer with 1 node, sigmoid activation, outputting only 1 value for class prediction.

The simple architecture of the CNN allowed for a model with as little as 0.3 million trainable parameters, much lesser than any of the pre-trained models used. This helped reduce both the structural and computational complexity of the model, allowing for faster results with minimal resource exhaustion. Similar to all previous models, the custom CNN was compiled using Adam optimiser, binary-crossentropy as loss function and accuracy as the metric. After much trial-and-error, the best average results were yielded after training it for 12 epochs with 100 steps per epoch, repeating for 5 times.

5.3 Model Evaluation

In order to evaluate the models, the research made use of a lot of metrics. Before moving on to analysing the results of the models, it is, first, important to know about these metrics to help understand and interpret the analysis.

5.3.1 Confusion Matrix

Confusion matrix, also known as error matrix, is a table layout that provides a visual representation or a summary of the performance of a classification or supervised model. In a binary classification, therefore there are 4 outcomes as below:

1. **True Positive (TP):** when both the predicted and actual outcomes are positive.
2. **True Negative (TN):** when both the predicted and actual outcomes are negative.
3. **False Positive (FP):** also called Type 1 error, is when the predicted outcome is positive but actual outcome is negative.
4. **False Negative (FN):** also called Type 2 error, is when the predicted outcome is negative but actual outcome is positive.

Confusion matrix is quite needed for measuring sensitivity, specificity, accuracy and AUC-ROC curves.

5.3.2 Accuracy

Accuracy is a method used for measuring the performance of a classification model and is usually expressed in percentage. It is the proportion of predictions where the predicted value is equal to the true value. As such the formula used to find this is:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5.1)$$

5.3.3 Loss

Loss or cost function is a measurement of how poorly or well behaved a model is after each iteration of optimisation. In other words, it takes account of the probability or uncertainty of a prediction based on how much it varied from its true value. Rather than using percentages, it is the summation of such errors made for each sample in training and validation sets.

The loss function used for the models is binary-crossentropy which compares each of the predicted probabilities to actual class output which can be either 0 or 1. It then calculates the score that penalises the probabilities based on the distance from the expected value. It is one of the most commonly used loss functions for binary classification problems.

5.3.4 Training vs Validation Accuracy and Loss

The accuracy and loss metrics are calculated both during training and validating the models. Training accuracy and validation accuracy are the accuracy values got after running the model through the training sets and validation sets respectively. Training accuracy shows how well the model is training whereas validation accuracy shows its ability to adapt to new dataset or, in other words, the quality of the model. When both the accuracy are close to equal, the model shows no overfitting, but as the training accuracy gets significantly higher than validation accuracy it shows signs of overfitting to training datasets, which means it is fitting to unnecessary noise while training.

Similarly, training and validation losses help in determining the quality and training of a model - that is, if it is overfit or underfit, or neither. Therefore, when training loss is similar to validation loss the model is just about right. Otherwise, if the training loss is quite lesser than validation loss it is overfitting, and vice versa for underfitting.

5.3.5 Sensitivity and Specificity

Sensitivity is the measure of proportions of actual positive cases that got predicted as positive, and is calculated using the formula:

$$sensitivity = \frac{TP}{TP + FN} \quad (5.2)$$

On the other hand, specificity is the measure of proportions of actual negative cases that got predicted as negative, and is calculated using the formula:

$$specificity = \frac{TN}{TN + FP} \quad (5.3)$$

It is a model's target to achieve both high sensitivity as well as high specificity.

5.3.6 Precision, Recall and F_1 Score

Precision is the ratio of actual positive cases to the total predicted positive cases, and is calculated using the formula:

$$precision = \frac{TP}{TP + FP} \quad (5.4)$$

Recall, on the other hand, is an alias for sensitivity and means the same thing. F_1 score, however, is the harmonic mean between precision and recall, and is particularly important for uneven class distributions. A good F_1 score (~ 1) indicates low FP and FN, meaning that both real threats are correctly identified and there are less false alarms. It is calculated using the formula:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (5.5)$$

5.3.7 ROC Curve and AUC

ROC curve is a probability curve for evaluating the model's performance for classification problems at various classification threshold settings. It is plotted using the values of sensitivity (or true positive rate (TPR)) and 1-specificity (or false positive rate (FPR)) on the y-axis and x-axis respectively. A sample ROC curve is shown in Figure 5.6 below.

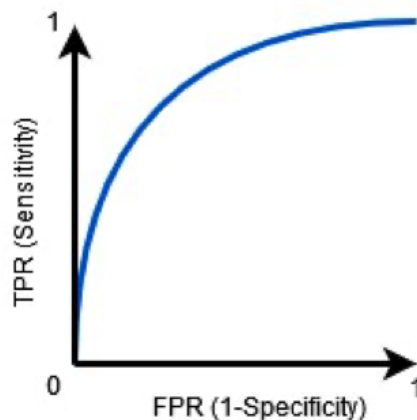


Figure 5.6: Sample ROC Curve

AUC is a degree of measure of the probability or ROC to evaluate the separability of the classifier between classes, and is the area under the ROC curve. Having an AUC value close to 1 simply refers to its ability to predict or separate true positives and true negatives successfully, while having AUC value close to 0 means that the model is predicting most of the true positives as negative and true negatives as positive, meaning it is still separating successfully but using inverted labels.

Features that give AUC merits are scale-invariant and classification-threshold-invariant, because it measures how well the predictions are ranked, and its quality without concerning the classification threshold that is chosen.

Chapter 6

Result Analysis

After training and testing all the models, it is now necessary to analyse the results obtained from them in order to distinguish between good and bad classifiers of breast cancer from raw mammograms. In order to have a general idea about the performance of the models, their test accuracy and loss, and average training time per epoch were evaluated, and can be seen in Table 6.1 below.

Model	Test Accuracy	Test Loss	Avg. Training Time Per Epoch (s)
ResNet50	0.9293	0.2864	19.0
MobileNet-V3 Small	0.9259	0.2662	3.0
VGG19	0.9293	0.3680	20.0
Custom CNN	0.9362	0.2017	1.0

Table 6.1: Test Accuracy, Loss and Training Times of the Models

Test accuracy depicts the models' ability to differentiate between healthy cells and cancer cells from unseen data. If the test accuracy is high, then it means that the model can distinguish between the cells much accurately. As can be seen from Table 6.1, the maximum test accuracy of 0.9362 was achieved by the custom CNN, while the least test accuracy of 0.9259 was scored by MobileNet-V3 Small, although, the difference is only around 1%. The accuracies of the other two models were between these values.

Test loss shows the variance between the actual and predicted results, where a low loss value indicates a more confident classifier. Here too, the custom CNN outperformed the other models with the least loss value of 0.2017. The maximum loss score of 0.3680 was achieved by VGG19, which is significantly higher than MobileNet-V3 Small and ResNet50 which scored 0.2662 and 0.2864 respectively.

Finally, the training time per epoch of the models were evaluated, which would allow to gain an insight into the complexity of the models, and their computational costs. Training time differs from machine to machine as it depends on the hardware configuration. The models of this research were trained on Google Colaboratory using GPU. The maximum time it took to complete an epoch was 20 seconds, which

was happening while training VGG19 model. Note that this is the average training time per epoch in seconds. Similarly, the ResNet50 model took 19 seconds. On the contrary, MobileNet-V3 Small and the custom CNN took much less time to train with only 3 seconds and 1 second per epoch respectively.

Apart from the aforementioned general evaluation metrics, one of the most widely used metric for binary classification problems is the AUC of the ROC curves of the models, which helps in separating the signal from the noise.

An AUC value between 0.5 and 1 is sought after as this means the model is able to produce a greater number of true positives and true negatives than false positives and false negatives. So, the higher the AUC value for a classifier, the better its ability to distinguish between positive and negative classes. Figure 6.1 below shows all the ROC-AUC of the models used in this research.

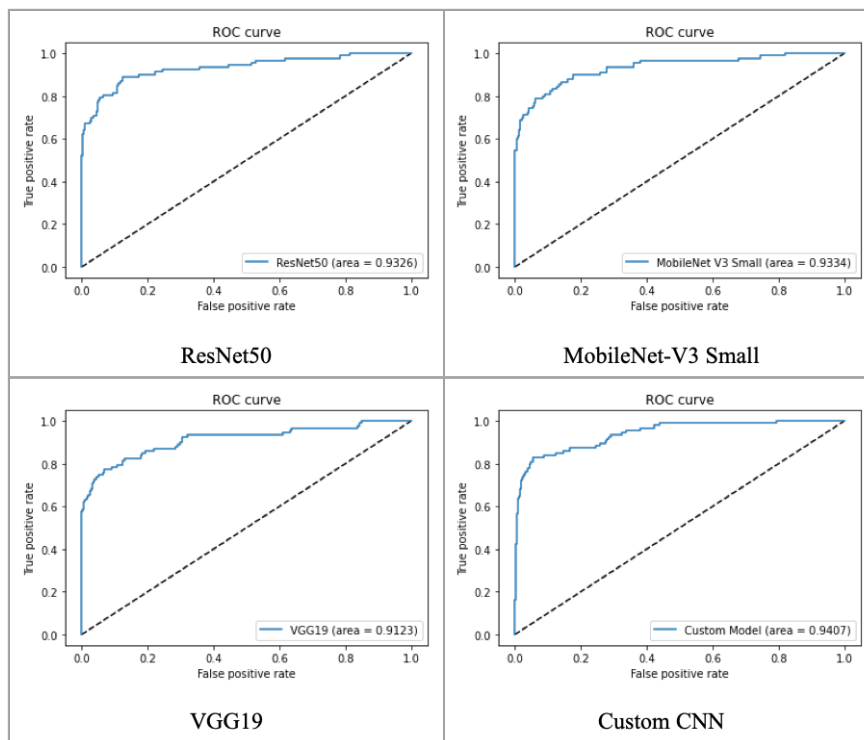


Figure 6.1: ROC-AUC of the Models

From Figure 6.1, it can be observed that the custom CNN resulted in the highest AUC with a value of 0.9407 while the VGG19 model had an AUC of 0.9123 which was the lowest out of the 4 models. Alternatively, both ResNet50 and MobileNet-V3 Small models produced similar results of 0.9326 and 0.9334 respectively.

For the higher thresholds in the graph, the custom CNN consistently had a higher true positive rate compared to the other models, meaning the model was able to correctly distinguish between benign and malignant with better confidence. However, the custom CNN did produce a lower true positive rate in the lower thresholds compared to the ResNet50 and MobileNet-V3 Small model while the VGG19 has a

similar value.

Nonetheless, it can be seen that all of the models have an AUC value higher than 0.90 which means that these models can correctly predict the onset of breast cancer with a higher probability compared to the predictions made by the current medical professionals in the field.

As seen till now, the models have produced splendid results in terms of accuracy, loss and AUC. However, these alone cannot guarantee that the models were good classifiers, as such results can also raise concerns of overfitting. Overfitting causes poor results when the same models are applied on unseen data, and hence is undesired. In order to check whether the models have overfitted, their training accuracy and loss were compared with their validation counterparts, and can be seen from Figure 6.2 below.

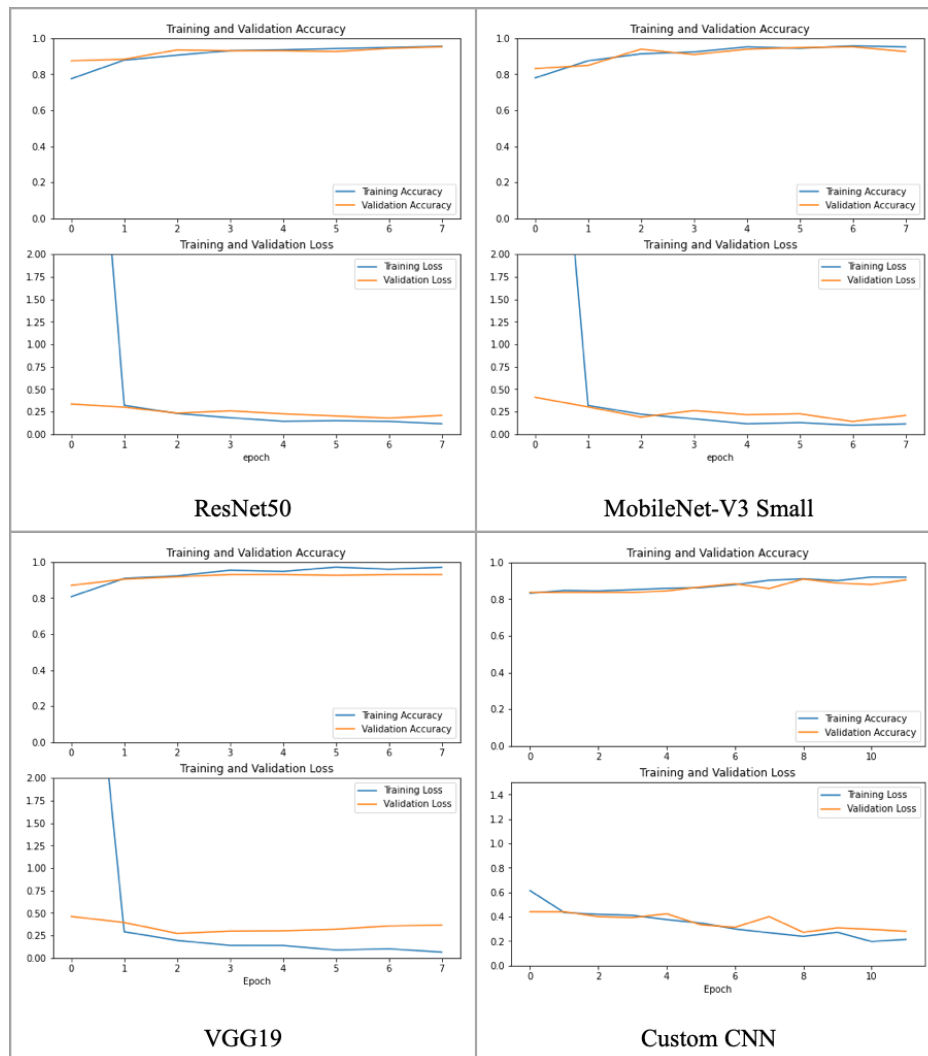


Figure 6.2: Training Vs Validation Metrics of the Models

Overfitting from the accuracy graph is found by checking the difference in training accuracy and validation accuracy. If the training accuracy is significantly higher

than validation accuracy then it is a case of overfitting. If the accuracy graphs of ResNet50 and MobileNet-V3 Small are checked, it can be seen that the validation and training accuracies are overlapping over each other, showing no hint of overfitting. The same observations can be made for the custom CNN as well, although there was a slight difference in one of the epochs, which was only present in one run and so, can be ignored as a case of anomaly. However, the VGG19's evaluation shows increasing difference in its training and validation accuracies, suggesting an overfit model. In order to confirm the observations, the loss graphs need to be evaluated as well.

Contrary to accuracy graph, in the loss graph, if training loss is significantly lower than validation loss then the model could be a case of overfitting. The loss graphs of ResNet50 and MobileNet-V3 Small show slight differences in their training and validation data-points, although they are quite negligible. On the other hand, the difference between training loss and validation loss for VGG19 is quite visible and significant. Hence, it confirms the initial analysis of being an overfit model, and cannot be considered a good CAD model for this research even after having admirable accuracy, loss and AUC. Finally, as for the custom CNN, the loss graph have overlapped well except at one epoch where the difference is quite high, but after that the graph again went back to being almost overlapping with only slight negligible differences. However, the accuracy graph too showed a disturbance at this exact epoch for one run, and hence, ensures the case of anomaly at that epoch for both accuracy and loss graphs.

From the observations, it can be deduced that except for VGG19, all the other models were able to classify breast cancer patients effectively, even when tested on unseen data.

As stated previously, accuracy and AUC are good measures of a model's performance when there is an even class distribution in the dataset, as well as when false negative and false positive predictions have the same cost. However, the MIAS dataset used in this research has an uneven class distribution, with more than 5 times as many non-cancerous tissues as cancerous ones. Moreover, false negative predictions have more cost as a genuine breast cancer threat might get neglected, compared to false positives where the alarmed patients can get a second test done to confirm their diagnosis. Hence, the models need to be evaluated based on advanced metrics.

The further metrics used for this research are average precision, recall and F_1 scores, and these scores for the models have been provided below in Table 6.2, calculated after 5 consecutive runs.

As can be seen from Table 6.2, all the models have much higher precision than they have recall, with the most disparity seen in MobileNet-V3 model. A higher precision indicates that there will be less false alarms among the healthy patients, whereas a poor recall score implies that most patients with breast cancer would be mislabelled as healthy which is even more menacing as the wrong diagnosis could prove to be fatal.

Model	Avg. Precision	Avg. Recall	Avg. F₁ Score
ResNet50	0.8983	0.7998	0.8462
MobileNet-V3 Small	0.9308	0.6544	0.7685
VGG19	0.8560	0.6586	0.7444
Custom CNN	0.9200	0.8025	0.8572

Table 6.2: Average Precision, Recall and F₁ Scores of the Models

MobileNet-V3 has the highest average precision of 0.9308, however, it also has the lowest recall of 0.6544 among all the models. Having said that, VGG19 has the poorest overall performance relative to others, with both low precision and recall. Among the pre-trained models, ResNet50 had the best overall performance, having high precision and recall scores. However, this time too, the custom CNN came out on top in terms of overall performance with an average precision score of 0.9200 and recall of 0.8025. This means that 92% of the patients predicted to have breast cancer by the model actually had the illness, while 80% of the patients who had breast cancer were correctly identified by the model.

Howbeit, there is a certain tradeoff between precision and recall, because of which the F₁ score is taken into account which is the harmonic mean of the aforementioned metrics. Hence, a high F₁ score would evince a better classifier. As predicted, the custom CNN had the best average F₁ score of 0.8572 due to its high precision and recall, with ResNet50, MobileNet-V3 Small and VGG19 coming at second, third and fourth positions having F₁ scores of 0.8462, 0.7685 and 0.7444 respectively.

To sum up the results evaluation, it can be safely said that the custom CNN proved to be the best CAD model for predicting breast cancer among patients using raw mammograms, having excelled at all the evaluation metrics, and that too after having the least number of trainable parameters among all the models used. It is simple, fast and a very accurate classifier to be used to aid radiologists in identifying breast cancer. Nonetheless, all the models mentioned in this paper have surpassed the metrics of [35], the paper that was used as the comparison base for this research.

Chapter 7

Discussion

After building and training several CNN models, the custom CNN model came out to be superior, excelling in all the metrics that it was analysed upon while trying to predict breast cancer from raw mammograms of patients. Not only did it have high accuracy and low loss, it also scored high in precision, recall and F_1 , all while ensuring that it does not overfit on the training data. However, it does raise a few questions as to how the model achieved this feat, reigning over the already established, industry-level pre-trained models. Although it cannot be said for certain what those reasons are without thorough inspection and scrutinising over the model, there are two in particular which can be inferred from how the research underwent.

The first probable reason could be related to the number of parameters of the models used. The pre-trained models were built for the ImageNet dataset, a database containing more than 14 million images organised into almost 22000 categories, which are mostly real-world objects such as animals, plants, home appliances etc. Hence, in order to segregate and classify test samples from such a large pool of different categorical images, the models used millions of trainable parameters, which looked for unique patterns that set each sample apart from the rest. However, the research at hand is only interested in the proper classification of breast cancer from raw mammograms. For such a naive task, the use of such advanced models proved to be overwhelming, performing poorly than anticipated. On the other hand, the custom CNN, which was built for this particular task, achieved better result with only about 300,000 parameters, eliminating the risks of overfitting on the training data. It is also to be noted that the model was developed through trial-and-error to find the suitable hyperparameters that produced the best results, with little regard to standard architectural design patterns followed by the developers of the pre-trained models.

Another probable reason could be associated with the type of layers used in the custom CNN model. Most pre-trained models, built specifically for the ImageNet dataset, utilises the Max Pooling layer, after the convolution layers, to choose the highest pixel value from each patch of each feature map, giving more importance to brighter and sharper pixels and edges when downsampling the images. It assumes some particular small features decide the class in which an image belong to, instead of analysing the whole image. However, in the case of mammograms, the final prediction cannot be reached by solely depending on some discrete features in

the image. Rather, patches of features distributed throughout the image decide its category, which are majorly greyish instead of being too dark or too bright. Because of this unique feature of mammograms, the Average Pooling layer was used as the first pooling layer in the custom CNN. This unconventional choice of the pooling layer could have positively influenced its results, making it a better classifier than the other pre-trained models.

Anyhow, the credit for the nobility of this research goes to the commendable approaches taken. First, the research dealt with the small dataset of MIAS. Since small datasets are more prone to bad classifiers and overfitting, the images extracted from MIAS were augmented, and the dataset size was increased by 9 times. Such a large dataset allowed the models to train better and effectively.

Second, the research incorporated the analysis of the whole raw mammograms instead of specific ROIs detected by radiologists beforehand. It allowed the models to mimic the actions of the radiologists in trying to detect breast cancer from the mammograms without relying on annotations which could have been erroneous.

Lastly, the research not only showed the efficiency of the existing pre-trained models in detecting breast cancer, but also had the objective of building a custom CNN for this task in particular. The custom model did not disappoint as it went ahead of all the pre-trained models in terms of the metrics they were evaluated upon.

Although it might seem that the research and its results were just perfect, it too had its fair share of limitations. The first problem with this research is the size of the original dataset. The dataset being small meant that data augmentation had to be applied in order to increase the size of the dataset. However, some of the augmented images had gone through too many changes to a point where they faintly resemble the original images. This also means that the data bias in the augmented images can be quite different from the original one resulting in suboptimal performance results from the model.

The other limitation is that the CNN model was trained and tested using only one dataset due to the unavailability of open-source datasets. As such, it is quite unsure if the model will work accurately on other datasets or if the model is only tailored to the specific dataset used, and as a result, the probability of the model providing accurate predictions for other datasets is unknown.

This research is just a stepping stone in the field of breast cancer prediction models, and opens up doors for further improvements going forward. Firstly, the CNN model developed in this research needs to be more generalised, by incorporating more datasets. Hence, more and more organisations in that regard should make their datasets public so that they can be used in research. Next, Deep Convolutional Generative Adversarial Network (DCGAN), a special type of ANN that can produce fake samples resembling the original dataset, could be used to generate more real-like samples instead of hard-augmentation manually that seems to distort the images. This would help in increasing the dataset size without compromising on quality. Lastly, in the future, better machine learning models could be developed that can

analyse and provide better imaging techniques from combining the various available now which can produce more sophisticated images, increasing the models' prediction capabilities.

Chapter 8

Conclusion

Breast cancer, a disease so deadly, that it still is one of the major causes of deaths among women worldwide. Howbeit, its diagnosis during the early stages can significantly reduce the mortality that it causes. In recent times, the development of automated medical applications and computer-aided diagnostic models has spurred the remarkable success of machine learning, especially deep learning and image processing. There is no one unique abnormality that act as the deciding factor in detecting breast cancer, but can be anything from masses and lumps to carcinomas, calcification or even asymmetry. Due to such varied possibilities of indicators of the disease, together with the dense breast tissues in some patients, humans alone cannot be expected to accurately identify the malignant tissues from raw mammograms. Hence, this paper focused on the use of deep convolutional neural networks to help the radiologist in reading a breast image.

First, the mammograms from mini-MIAS database were extracted and passed through several pre-processing and augmentation steps to enhance clarity of the images and increase the size of the dataset. Although the research shed light on the efficiency of the existing pre-trained CNN models in classifying breast cancer, the main objective was to build and test a CNN model that has been custom-made for this task only. The enhanced CNN that was formulated obtained high test accuracy, AUC, precision, recall and F_1 scores while having minimal to no overfitting when applied to the dataset.

Even then, CAD models need to undergo further improvements before they can be used as standalone detectors of breast cancer. Nonetheless, at its current stage, such models have the potential to aid the radiologists make more informed predictions about the patients from raw mammograms. It is only a matter of time before ML can be used to detect the abnormality in the breast early, thus, determining the most appropriate subsequent treatment to be administered to the patient to help bring down the number of deaths caused by breast cancer.

Bibliography

- [1] D. H. Hubel and T. N. Wiesel, “Receptive fields and functional architecture of monkey striate cortex,” *The Journal of Physiology*, vol. 195, no. 1, pp. 215–243, 1968. DOI: 10.1113/jphysiol.1968.sp008455.
- [2] K. Fukushima, “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position,” *Biological Cybernetics*, vol. 36, no. 4, pp. 193–202, Apr. 1980. DOI: 10.1007/BF00344251.
- [3] W. H. Wolberg, W. N. Street, and O. L. Mangasarian, “Breast cancer wisconsin (diagnostic) data set,” *UCI Machine Learning Repository*, 1992. [Online]. Available: [https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+\(original\)](https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)).
- [4] J. Suckling, “The mammographic image analysis society digital mammogram database,” *Excerpta Medica International Congress*, vol. 1069, pp. 375–378, 1994. [Online]. Available: <http://peipa.essex.ac.uk/info/mias.html>.
- [5] H. Hamilton, N. Shan, and N. Cercone, “Riac: A rule induction algorithm based on approximate classification,” *International Conference on Engineering Applications of Neural Networks*, University of Regina, May 1996.
- [6] B. Šter and A. Dobnikar, “Neural networks in medical diagnosis: Comparison with other methods,” *Proceedings of the International Conference on Engineering Applications of Neural Networks*, pp. 427–430, Jan. 1996.
- [7] K. Bennett and J. Blue, “A support vector machine approach to decision trees,” in *1998 IEEE International Joint Conference on Neural Networks Proceedings*, vol. 3, IEEE World Congress on Computational Intelligence, 1998, pp. 2396–2401. DOI: 10.1109/IJCNN.1998.687237.
- [8] G. Bradski, “The opencv library,” *Dr. Dobb’s Journal of Software Tools*, 2000.
- [9] R. Setiono, “Generating concise and accurate classification rules for breast cancer diagnosis,” *Artificial Intelligence in Medicine*, vol. 18, no. 3, pp. 205–219, Apr. 2000. DOI: 10.1016/S0933-3657(99)00041-X.
- [10] D. Goodman, L. Boggess, and A. Watkins, “Artificial immune system classification of multiple-class problems,” *Proceedings of the Artificial Neural Networks in Engineering*, vol. 12, pp. 179–183, Jan. 2004.
- [11] D. of Radiology, “Wisconsin state cancer reporting system,” *University of Wisconsin, Madison*, 2004.
- [12] L. Hadjiiski, B. Sahiner, M. A. Helvie, *et al.*, “Breast masses: Computer-aided diagnosis with serial mammograms,” *Radiology*, vol. 240, no. 2, pp. 343–356, 2006, PMID: 16801362. DOI: 10.1148/radiol.2401042099.

- [13] H. Alhammady, “Weighted naive bayesian classifier,” IEEE/ACS International Conference on Computer Systems and Applications, Jun. 2007, pp. 437–441, ISBN: 1-4244-1031-2. DOI: 10.1109/AICCSA.2007.370918.
- [14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.
- [15] T. Ayer, O. Alagoz, J. Chhatwal, J. Shavlik, C. Kahn Jr, and E. Burnside, “Breast cancer risk evaluation with artificial neural networks revisited,” *Cancer*, vol. 116, no. 14, pp. 3310–3321, Jul. 2010. DOI: 10.1002/cncr.25081.
- [16] T. Balakumaran, I. Vennila, and C. G. Shankar, “Detection of microcalcification in mammograms using wavelet transform and fuzzy shell clustering,” *International Journal of Computer Science and Information Technology*, vol. 7, no. 1, pp. 121–125, 2010. arXiv: 1002.2182. [Online]. Available: <http://arxiv.org/abs/1002.2182>.
- [17] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, “Scikit-learn: Machine learning in python,” *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011. [Online]. Available: <http://jmlr.org/papers/v12/pedregosa11a.html>.
- [18] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” *CoRR*, vol. abs/1207.0580, 2012. [Online]. Available: <http://arxiv.org/abs/1207.0580>.
- [19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems*, vol. 25, Curran Associates, Inc., 2012, pp. 1097–1105.
- [20] I. Saritas, “Prediction of breast cancer using artificial neural networks,” *Journal of Medical Systems*, vol. 36, no. 5, pp. 2901–2907, Aug. 2012. DOI: 10.1007/s10916-011-9768-0.
- [21] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed., ser. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc, Mar. 2013, ISBN: 9781118548387. DOI: 10.1002/9781118548387.
- [22] V. Dheeba, N. Singh, and A. Singh, “Breast cancer diagnosis: An intelligent detection system using wavelet neural network,” in *Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2013*, vol. 247, Springer, Cham, Jan. 2014, pp. 111–118, ISBN: 978-3-319-02930-6. DOI: 10.1007/978-3-319-02931-3_14.
- [23] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *International Conference on Learning Representations*, Dec. 2014. [Online]. Available: <https://arxiv.org/abs/1412.6980>.
- [24] M. Abadi, A. Agarwal, P. Barham, *et al.* “TensorFlow: Large-scale machine learning on heterogeneous systems.” (2015), [Online]. Available: <https://www.tensorflow.org/>.

- [25] M. Abadi, A. Agarwal, P. Barham, *et al.*, “Tensorflow: Large-scale machine learning on heterogeneous distributed systems,” *TensorFlow*, 2015. [Online]. Available: <http://download.tensorflow.org/paper/whitepaper2015.pdf>.
- [26] B. M. Gayathri and C. P. Sumathi, “Mamdani fuzzy inference system for breast cancer risk detection,” in *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCCIC)*, 2015, pp. 1–6. DOI: 10.1109/ICCCIC.2015.7435670.
- [27] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 770–778. [Online]. Available: <http://arxiv.org/abs/1512.03385>.
- [28] M. Karabatak, “A new classifier for breast cancer detection based on naive bayesian,” *Measurement*, vol. 72, pp. 32–36, Aug. 2015. DOI: 10.1016/j.measurement.2015.04.028.
- [29] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: <http://arxiv.org/abs/1409.1556>.
- [30] C. Szegedy, W. Liu, Y. Jia, *et al.*, “Going deeper with convolutions,” in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9. DOI: 10.1109/CVPR.2015.7298594.
- [31] “What is cancer?” *National Cancer Institute*, 2015. [Online]. Available: <https://www.cancer.gov/about-cancer/understanding/what-is-cancer>.
- [32] B. M. Gayathri and C. P. Sumathi, “Comparative study of relevance vector machine with various machine learning techniques used for detecting breast cancer,” in *2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCCIC)*, 2016, pp. 1–5. DOI: 10.1109/ICCCIC.2016.7919576.
- [33] D. Lévy and A. Jain, “Breast mass classification from mammograms using deep convolutional neural networks,” *CoRR*, vol. abs/1612.00542, 2016. arXiv: 1612.00542. [Online]. Available: <http://arxiv.org/abs/1612.00542>.
- [34] Sentdex, *Full classification example with convnet*, 2016. [Online]. Available: <https://www.kaggle.com/sentdex/full-classification-example-with-convnet>.
- [35] Y. J. Tan, K. S. Sim, and F. F. Ting, “Breast cancer detection using convolutional neural networks for mammogram imaging system,” in *2017 International Conference on Robotics, Automation and Sciences (ICORAS)*, IEEE Xplore, 2017, pp. 1–5. DOI: 10.1109/ICORAS.2017.8308076.
- [36] P. H. Tsui, M. M. Mehdy, P. Y. Ng, E. F. Shair, N. I. M. Saleh, and C. Gomes, “Artificial neural networks in image processing for early detection of breast cancer,” *Computational and Mathematical Methods in Medicine*, vol. 2017, p. 15, 2017. DOI: 10.1155/2017/2610628. [Online]. Available: <https://doi.org/10.1155/2017/2610628>.

- [37] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, “Random erasing data augmentation,” *CoRR*, vol. abs/1708.04896, 2017. [Online]. Available: <http://arxiv.org/abs/1708.04896>.
- [38] S. Aboutalib, A. Abdelrahim, W. Berg, M. Zuley, J. Sumkin, and S. Wu, “Deep learning to distinguish recalled but benign mammography images in breast cancer screening,” *Clinical Cancer Research*, vol. 24, Oct. 2018. DOI: 10.1158/1078-0432.CCR-18-1115.
- [39] M. Amrane, S. Oukid, I. Gagaoua, and T. EnsarĪ, “Breast cancer classification using machine learning,” in *2018 Electric, Electronics, Computer Science, Biomedical Engineerings’ Meeting (EBBT)*, 2018, pp. 1–4. DOI: 10.1109/EBBT.2018.8391453.
- [40] Y. Guo, L. Shuhui, Z. Li, and X. Shang, “Bcdforest: A boosting cascade deep forest model towards the classification of cancer subtypes based on gene expression data,” *BMC Bioinformatics*, vol. 19, Apr. 2018. DOI: 10.1186/s12859-018-2095-4.
- [41] E. K. Kim, H. E. Kim, K. Han, *et al.*, “Applying data-driven imaging biomarker in mammography for breast cancer screening: Preliminary study,” *Scientific Reports*, vol. 8, p. 2762, Feb. 2018. DOI: 10.1038/s41598-018-21215-1.
- [42] S. Mambou, O. Krejcar, P. Maresova, A. Selamat, and K. Kuca, “Breast cancer detection using infrared thermal imaging and a deep learning model,” *Sensors*, vol. 18, no. 9, p. 2799, Aug. 2018. DOI: 10.3390/s18092799.
- [43] S. Sharma, A. Aggarwal, and T. Choudhury, “Breast cancer detection using machine learning algorithms,” *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, pp. 114–118, 2018. DOI: 10.1109/ctems.2018.8769187.
- [44] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, “Convolutional neural networks: An overview and application in radiology,” *Insights into Imaging*, vol. 9, no. 4, pp. 611–629, 2018. DOI: 10.1007/s13244-018-0639-9.
- [45] “Breast cancer early detection and diagnosis,” *American Cancer Society*, 2019. [Online]. Available: <https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection.html>.
- [46] N. Harbeck, F. Penault-Llorca, J. Cortes, *et al.*, “Breast cancer,” *Nature Reviews Disease Primers*, vol. 5, no. 1, p. 66, 2019. DOI: 10.1038/s41572-019-0111-2.
- [47] J. Jaworek-Korjakowska, P. Kleczek, and M. Gorgon, “Melanoma thickness prediction based on convolutional neural network with vgg-19 model transfer learning,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 2748–2756. DOI: 10.1109/CVPRW.2019.00333.
- [48] “Limitations of mammograms,” *American Cancer Society*, 2019. [Online]. Available: <https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/mammograms/limitations-of-mammograms.html/>.
- [49] Lixinso, “Mobilenet review,” *Medium*, 2019. [Online]. Available: <https://lixinso.medium.com/mobilenet-c08928f2dba7>.

- [50] J. Rodriguez, “Google open sources mobilenetv3 with new ideas to improve mobile computer vision models,” *KDnuggets*, Dec. 2019. [Online]. Available: <https://www.kdnuggets.com/2019/12/google-open-sources-mobilenetv3-improve-mobile-computer-vision-models.html>.
- [51] A. Rodriguez-Ruiz, E. Krupinski, J.-J. Mordang, *et al.*, “Detection of breast cancer with mammography: Effect of an artificial intelligence support system,” *Radiology*, vol. 290, no. 2, pp. 305–314, 2019, PMID: 30457482. DOI: 10.1148/radiol.2018181371. eprint: <https://doi.org/10.1148/radiol.2018181371>.
- [52] L. Shen, L. Margolies, J. Rothstein, E. Fluder, R. McBride, and W. Sieh, “Deep learning to improve breast cancer detection on screening mammography,” *Scientific Reports*, vol. 9, pp. 1–12, Aug. 2019. DOI: 10.1038/s41598-019-48995-4.
- [53] I. Aryan, “Clahe histogram equalization – opencv,” *GeeksforGeeks*, 2020. [Online]. Available: <https://www.geeksforgeeks.org/clahe-histogram-equalization-opencv/>.
- [54] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, “Albumentations: Fast and flexible image augmentations,” *Information*, vol. 11, no. 2, 2020, ISSN: 2078-2489. DOI: 10.3390/info11020125.
- [55] J. Ferlay, M. Ervik, F. Lam, *et al.*, “Global cancer observatory: Cancer today,” *Lyon, France: International Agency for Research on Cancer*, 2020. [Online]. Available: <https://gco.iarc.fr/today>.
- [56] F. Bray, M. Laversanne, E. Weiderpass, and I. Soerjomataram, “The ever-increasing importance of cancer as a leading cause of premature death worldwide,” *Cancer*, vol. 127, 2021.
- [57] H. Sung, J. Ferlay, R. L. Siegel, *et al.*, “Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021. DOI: <https://doi.org/10.3322/caac.21660>. eprint: <https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21660>. [Online]. Available: <https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21660>.
- [58] “Cancer,” *World Health Organization, WHO*, [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/cancer>.
- [59] A. Kaushik, “Understanding the vgg19 architecture,” *OpenGenus Foundation*, [Online]. Available: <https://iq.opengenus.org/vgg19-architecture/>.
- [60] A. Kaushik, “Understanding resnet50 architecture,” *OpenGenus Foundation*, [Online]. Available: <https://iq.opengenus.org/resnet50-architecture/>.