Intelligent Computer-Aided Systems for Breast Cancer Classification in Digital Breast Tomosynthesis Scans

Alaa M. Adel El-Shazli

A thesis submitted in partial fulfilment of the requirements of University of Staffordshire for the degree of Doctor of Philosophy

University of Staffordshire
School of Digital, Technology, Innovation and Business
United Kingdom

Abstract

Breast cancer is a significant worldwide health concern, with high incidence and fatality rates. Early identification is critical to improving patient outcomes and reducing the overall burden of the disease. This thesis contributes to knowledge by developing five novel systems for multi-class classification of Digital Breast Tomosynthesis (DBT) scans as normal, benign, or malignant.

This thesis presents unique methodologies and combines them to create five systems. The first system, DeepEval System (DE System), compares six cutting-edge DL models for feature extraction prior to classification using a Support Vector Machine (SVM), serving as a benchmarking tool.

The Mod_AlexNet System (MA System) is presented thereafter. It is a novel system that modifies the traditional AlexNet by incorporating max-pooling layers and batch normalisation layers. These modifications are designed to improve the classification performance. Various optimizers were tested and compared while training Mod_AlexNet with different batch sizes to optimize the performance. In the Feature Fusion and Selection with Ensemble Classifier (FFS-EC) System, feature fusion is integrated, followed by several feature selection models and a majority voting ensemble classification model.

The Multi-Head Mod_AlexNet Attention (MHMA) system introduces a novel attention model to the previously developed Mod_AlexNet. This attention framework focuses on the most relevant parts of the input images, improving the ability of the system to represent important features and thereby enhancing the overall classification performance. Finally, the Hybrid Multi-Head Self-Attention Model with Feature Fusion, Selection, and IVECM for Enhanced DBT Classification (HMSA-FFS-IVECM) System integrates Mod_AlexNet with the attention model, feature fusion, and selection models, as well as a newly developed Integrated Voting Ensemble Classification Model, IVECM, that incorporates class and classifier weights. This comprehensive integration maximizes the classification performance, particularly for the abnormal classes, benign and malignant, which are minorities in the dataset.

The systems were tested using a publicly available dataset, called Breast Cancer Screening – Digital Breast Tomosynthesis (BCS-DBT) dataset (Buda et al., 2020). The HMSA-IVECM System achieved a remarkable specificity of 62.20%, significantly outperforming: DE System (21.43%), MA System (23.91%), FFS-EC (43.07%), and MHMA System (51.99%). The proposed HMSA-IVECM system consistently outperforms existing methods across all scenarios. For benign versus malignant classification, it achieved the highest accuracy of 91.24% and 91.09% in both scenarios, surpassing the closest competitor (Farangis Sajadi Moghadam and Rashidi, 2023) by over 2.5% and Hassan et al. (2024) by over 6%. For normal versus abnormal and cancerous versus non-cancerous cases, HMSA-IVECM demonstrated superior accuracy (94.53% and 93.81%, respectively), showing substantial improvement in sensitivity, precision, and F1-score compared to prior models. These systems contribute to the development of automated breast cancer classification technologies, which promises to greatly enhance the early diagnosis and classification of breast abnormalities.

Acknowledgments

Though only my name appears on the cover of this thesis, many great people have contributed to this research. I am indebted to everyone for their support and guidance throughout my pursuit of this thesis, and my sincere apologies to those I may have forgotten to mention.

I would like to express my gratitude to my supervisors, Prof. Abdel Hamid Soliman, Prof. Sherin M. Youssef, and Prof. Claude Chibelushi, whose expertise, patience, and understanding greatly enhanced my experience. Their guidance and support were invaluable at all levels of research.

I would also like to acknowledge and express my special gratitude and thanks to the previous Head of the Radiology Department and former Dean of Alexandria Faculty of Medicine, Prof. Dr. Yehia Haleem, and Professor of Radiodiagnosis, Prof. Adel Mohamed Rizk, for helping me understand more about breast scans and breast cancer.

Many friends have helped me through these difficult years. I greatly value their friendship and support.

Most importantly, none of this would have been possible without the love, patience, and support of my family. To my family, particularly my mom and dad, this thesis is dedicated to you. Your endless love, care, support, and strength have been a pillar throughout my life. Your unwavering faith in me, your sacrifices, and your encouragement have been the bedrock upon which this achievement stands. I am deeply grateful for your boundless love and for always being there for me through every challenge and triumph.

Contents

LIST OF FIGURES

LIST OF TABLES

List of Abbreviations

2D-DCNN	Two Dimensional Deep Convolutional Neural Networks
2FPI	2 False Positives per Image
3D-AlexNet	Three-Dimensional AlexNet
3D-DCNN	Three-Dimensional Deep Convolutional Neural Networks
3D-GMIC	Three-Dimensional Globally Aware Multiple Instance Classifier
3D-Mask	Three-Dimensional Mask
AI	Artificial intelligence
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
AUC	Area Under the Curve
BCE	Binary Cross Entropy
BCS-DBT	Breast Cancer Screening - Digital Breast Tomosynthesis
CAD	Computer-Aided Diagnosis
CLAHE	Contrast Limited Adaptive Histogram Equalisation
CLS	Classifier Weighting
CMMD	Chinese Mammography Database
CNN	Convolutional Neural Network
DBT	Digital Breast Tomosynthesis
DCNN	Deep Convolutional Neural Network
DDSM	Digital Database for Screening Mammography
DE	DeepEval
DL	Deep Learning
DM	Digital Mammography
DT	Decision Trees
DTL	Double Transfer Learning
EMMFFN	End-to-End Multi-scale Multi-level Features Fusion Network
FBP	Filtered Back Projection
FC	Fully Connected
FCN	Fully Connected Network
FDA	Fisher Discriminant Analysis
FETL	Feature Extraction-based Transfer Learning

FFDM	Full-Field Digital Mammography
FFS-EC	Feature Fusion and Selection with Ensemble Classifier
FN	False Negative
FP	False Positive
FPN	Feature Pyramid Network
GAP	Global Average Pooling
GCN	Graph Convolution Network
GLCM	Gray-Level Co-Occurrence Matrix
HM	Histogram Matching
HMSA-IVECM	Hybrid Multi-head Self-Attention with Integrated Voting Ensemble Classification Model
HOG	Histogram of Oriented Gradients
HSV	Hue, Saturation, and Value
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
IVECM	Integrated Voting Ensemble Classification Model
KNN	K-Nearest Neighbours
LLSTM	Longitudinal LSTM model
LRN	Local Response Normalisation
LSTM	Long Short-Term Memory
MA	Mod_AlexNet
MCs	Microcalcification Clusters
MGCN	Multi-scale Graph Convolution Network
MHMA	Multi-Head Mod_AlexNet Attention
MIL	Multiple Instance Learning
MIP	Maximum Intensity Projections
MLO	MedioLateral Oblique
MLP	Multi-Layer Perceptron
MLTL	Multi-Level Transfer Learning
MRI	Magnetic Resonance Imaging
mRMR	Minimum Redundancy Maximum Relevance
NB	Naive Bayes
PMV-MG	Proprietary Multi-Vendor - Mammography
PPV	Positive Predictive Value
RBF	Radial Basis Function

RBSM	Rank Based Score Modification
ReLU	Rectified Linear Unit
RMSProp	Root Mean Square Propagation
RNNs	Recurrent Neural Networks
ROI	Region of Interest
RPN	Region Proposal Network
SA	Self-Attention
SGD	Stochastic Gradient Descent
SGDM	Stochastic Gradient Descent with Momentum
STL	Single Transfer Learning
SVM	Support Vector Machines
TN	True Negative
TP	True Positive
UCHC	University of Connecticut Health Centre
UPMC	University Pierre and Marie CURIE
US	Ultrasound
VOI	Volume of Interest

Chapter 1 Introduction

1.1 Overview and Motivation

Worldwide, breast cancer is the most common cancer to be diagnosed, and in the last few years, its incidence has increased. In 2022, breast cancer represented the highest incidence rate among cancers affecting women globally, according to statistics from the World Health Organization (WHO), IARC statistics (Bray et al., 2024). It accounted for 23.8% of all new cancer diagnosis cases in women, with approximately 2.3 million cases. In contrast, lung cancer, the second most prevalent cancer among women, represented only 9% of new cases. Furthermore, breast cancer also led in cancer-related mortality, responsible for 670,000 deaths, representing 15% of all female cancer-related deaths in 2022. This surpasses lung cancer, which accounted for 580,000 deaths, or 13% of cancer mortalities in women for the same year (Bray et al., 2024).

From 1975 to 1989, the breast cancer-related mortality rate increased by 0.4% yearly; however, from 1989 to 2020, it decreased by 43%. The notable decline in breast cancer mortality has been attributed to advances in early identification by screening mammography, breast scans and improvements in personalised therapy (Giaquinto et al., 2022). Early detection and treatment dramatically improve the odds of survival for those with breast cancer. Depending on the stage of the cancer, there are significant differences in the survival rates; earlier diagnosis leads to more effective therapy.

Digital Breast Tomosynthesis (DBT) is an advanced imaging technique that generates three-dimensional (3D) images of the breast using low-dose X-rays. Unlike traditional mammography, which produces two-dimensional images, DBT captures multiple angles of the breast and reconstructs them into a detailed 3D image. This allows for better visualisation of breast tissue, improving the detection of small abnormalities and reducing false positives and recall rates. DBT scans have become an essential tool in modern breast cancer diagnosis, helping to detect malignancies more accurately, especially in women with dense breast tissue.

Breast calcifications, or tiny calcium deposits, can occur in women as they age. Although common, calcifications can be early signs of abnormal cell growth. The

size, shape, and distribution of calcifications can provide information on whether more testing is required to rule out malignancy (Logullo et al., 2022).

Artificial intelligence (AI) is a vital tool in modern life, both for improving and protecting people. Computer models known as Medical Computer-Aided Diagnosis (CAD) Systems improve the precision of medical image identification, diagnosis, and interpretation. Cutting-edge medical CAD algorithms make significant improvements to imaging medicine by helping doctors diagnose and treat patients quickly and effectively. CAD medical systems are used in surgical procedures like laparoscopy, to help radiologists identify anomalies in medical images, and to help interpret blood sample abnormalities and diseases (Guo et al., 2022). In the following section, the motivation for the thesis and the formulation of the problem statement will be addressed.

1.2 Problem Statement

Breast cancer is a widespread worldwide health concern that affects people worldwide. The reason for its widespread incidence is that it is one of the most common cancers in the world, with breast cells being the source of its origin. Early identification of breast cancer is a crucial aspect of healthcare that requires ongoing developments and research.

Breast lesion classification into benign, malignant, and normal classes is still a difficult task, even with the improved diagnostic capabilities provided by tomosynthesis scans. The existing diagnostic approaches have difficulty reaching high accuracy, particularly in multi-class classification across benign, malignant, and normal classes. Accurate classification is significantly limited by the small differences in size, shape, and texture found in abnormal cases, which make it much more difficult to distinguish between benign and malignant cases. The existing gap in the literature makes this issue further significant since earlier research mostly concentrated on binary classifications, such as normal vs. abnormal, benign vs. malignant, or cancerous vs. non-cancerous, ignoring the importance of developing multi-class classification systems. Such systems are crucial for differentiating between multiple types of abnormalities, enabling the classification of images not only as normal or abnormal but also distinguishing between benign and malignant abnormalities.

Accurately diagnosing breast cancer is severely restricted by the fundamental variability in breast structure. Individual differences exist in the size and density of breasts, which complicates the interpretation of images. These variations complicate the multi-class classification and challenge the development of effective multi-class CAD systems.

The challenges faced in the process of developing the contributions to knowledge, including the systems, reported in this thesis, are outlined in the following section, along with brief explanations of each challenge.

1.3 Challenges

Developing automated algorithms for tomosynthesis scan assessment is challenged by numerous factors. These challenges could result in a misdiagnosis that causes further unnecessary testing and x-ray exposure, or they could cause a failure to identify an existing abnormality that, if undetected, could have serious consequences for the patient. To successfully implement an intelligent computer-aided system for accurate classification of tomosynthesis scans, several challenges must be overcome. The key challenges and gaps include:

1. Multi-class classification accuracy

Previously developed systems faced difficulties in effectively differentiating between benign, normal, and malignant classes, often resulting in low accuracy that is insufficient for reliable clinical application. Misclassification rate has challenged the development of a robust diagnostic system. The limitation of low classification accuracy of these systems highlights the urgent need for novel strategies to improve the accuracy and reliability of breast cancer diagnosis.

2. Varying breast density and size

The presence of different breast sizes and densities adds another degree of complication, making distinguishing between benign, normal, and malignant cases challenging. Breast structure varies widely, which adds to the varied looks on scans and makes the classification task more complex. This challenge highlights the various aspects of breast cancer diagnosis, demanding enhanced techniques capable of efficiently going through the diverse characteristics offered by varying breast sizes and densities to make more accurate and reliable classifications.

3. Benign vs malignant classification

The challenge of distinguishing between benign and malignant abnormalities is further complicated by the sophisticated characteristics of these variations, making the accurate discrimination between these two classes hard. In order to enhance precision and reliability in identifying breast lesions, sophisticated models capable of identifying small differences are required due to the complexities of abnormalities, which are characterised by small variations in size, shape, and texture.

The subsequent section provides a detailed explanation of the aims and objectives of the thesis. These objectives act as an organised framework that guides the research studies and adds to the overall methodology and structure of the study.

1.4 Aims and Objectives

The research aims to create an intelligent three-class CAD system that accurately classifies breast tomosynthesis scans into normal, benign, or malignant categories. It aims to identify salient features of the image that can best discriminate between the three classes, regardless of breast density and size, while also extracting robust and compact features from the images to reliably distinguish between the classes. Additionally, the research aims to develop a classification model that maximises the between-class variance, and minimises the overlap in class distributions, to enhance the ability to differentiate between the classes. The parameters of the classification model will be optimised to maximise performance. The performance of the model will be evaluated on tomosynthesis scans to test its accuracy and reliability. The research aims to a clinical impact by improving breast cancer classification precision and reducing False Positive (FP) and False Negative (FN) rates.

The objectives of the thesis are to:

- Perform a comprehensive analysis of current machine learning and deep learning techniques, as well as state-of-the-art breast cancer classification models for breast scans.
- Examine cutting-edge feature extraction methods and their impact on a three-class classification model.
- Explore and utilise feature fusion and selection models to address challenges in multi-class classification.

- Develop an optimal classification model that is capable of handling non-linearity of class boundaries, and in which the classes are sufficiently distinct to allow for better separation among the three classes.
- Analyse the optimal classification approach for increasing classification performance across all metrics.
- Evaluate the performance of the CAD systems through experiments, comparing and validating the results with the findings from the literature review using appropriate performance metrics.
- Contribute to the academic field through the publication and sharing of research findings

1.5 Contributions to Knowledge

In order to distinguish between normal, benign, and malignant tomosynthesis scans, this thesis presents and explores five systems for the classification of breast tomosynthesis scans: one benchmarking system (DeepEval System – DE System) and four novel Deep Convolutional Neural Network (DCNN)-based systems. The following is a summary of the main contributions of the thesis:

1. The proposed Mod_AlexNet System, referred to as the MA System, represents a significant advancement in breast tomosynthesis classification.. The Mod_AlexNet architecture is specifically designed to extract and classify features from tomosynthesis scans. Mod_AlexNet modified the traditional AlexNet. To optimise performance, the model is thoroughly trained using a range of optimisers and batch sizes. A comparative evaluation is conducted to assess the performance of Mod_AlexNet against that of AlexNet, considering varying optimisers and batch sizes during training. The main goal of this architectural modification is to maximise the classification performance of the traditional AlexNet. Detailed insights into this system are elaborated upon in Chapter 7.
2. In the Feature Fusion and Selection with Ensemble Classifier System, named FFS-EC, an enhanced system for the classification of DBT data through the integration of DL models with feature fusion, selection and classification ensemble models is introduced. The system extracts features from both the deep learning models and the Histogram of Oriented Gradients (HOG) to

extract more prominent features. After extracting features, a series of fusion and selection processes are applied sequentially. The selected features are subsequently classified using several classifiers. The final model in this system, an ensemble classifier, combines the predictions of multiple classifiers. For training using the DBT dataset, Feature Fusion and Selection with Ensemble Classifier (FFS-EC), uses two pre-trained models, ResNet-50 and SqueezeNet, in addition to the previously developed Mod_AlexNet deep learning model. The integration of feature fusion, selection, and ensemble classifier reduces dimensionality and combines diverse predictive strengths to select the most significant features, which can effectively discriminate between different classes, resulting in improved classification performance across different measured performance measures. Chapter 8 provides detailed insights into the methodologies employed in FFS-EC.

3. One novel contribution of this thesis is the construction of a three-layer Multi-Head Mod_AlexNet Attention (MHMA) model that utilises the final pooling outputs of Mod_AlexNet. Specifically designed to tackle issues caused by differences in breast sizes in DBT data, this model combines Fully Connected (FC) layers and SA techniques. The adaptability of the model is improved by training it with SGDM, Adam, and RMSProp optimisers over a range of batch sizes. The model enhances adaptability and addresses challenges caused by differences in breast sizes, resulting in improved diagnostic accuracy, especially in the abnormal classes. In Chapter 9, the methodical processes used in the MHMA System are described in depth, emphasising how this novel approach greatly enhances classification performance by effectively addressing complexities related to varying breast sizes in tomosynthesis images.
4. The most significant contribution of this thesis comes from the Hybrid Multi-head Self-Attention with Integrated Voting Ensemble Classification Model (HMSA-IVECM) System, which is thoroughly explored in Chapter 10, and serves as an integrated solution addressing the challenges encountered in tomosynthesis classification. Developing an innovative ensemble classifier, IVECM, represents a key contribution of this thesis, with the objective of resolving the critical issue of class imbalance in the data set being utilised is a crucial component of this thesis. Weights for each classifier are assigned

carefully, with the weighted average F1-score and specificity for each classifier being considered when determining the classifier weights. Furthermore, class weights are generated and integrated with the classifier weights, guaranteeing an intelligent and efficient ensemble model. The HOG descriptors and features extracted from the Mod_AlexNet Self-attention (SA) module were carefully combined through concatenation, and several feature selection algorithms are utilised. The final forecast is then obtained by feeding the selected features from each selection model into the IVECM. This system integrates individual contributions and effectively addresses the challenges posed by the class imbalance in the dataset, resulting in more accurate and reliable diagnostic results. Chapter 10 provides details of structures of the HMSA-IVECM System and how crucial it is to overcome these challenges.

1.6 Thesis Organisation

There are eleven chapters in this thesis. A review of the medical background on breast structure is provided in Chapter 2, which also discusses the imaging modalities employed for breast cancer screening and presents a variety of findings from breast scans. Chapter 3 explores CAD systems and provides a thorough summary of the approaches utilised in the systems developed during this thesis. A thorough overview of the literature is provided in Chapter 4, which includes studies on tomosynthesis scans using various private datasets, and the dataset utilised in the work presented in this thesis. Chapter 5 outlines the employed dataset, experimental framework, and performance metrics for evaluating system efficacy. Chapters 6-10 emphasise the five developed systems, explaining their methodology, experimental setting, and presentation of the results.

Chapter 6 introduces a comparative evaluation system for feature extraction, which investigates six state-of-the art deep learning models and employs them as feature extraction models before classifying the features using an SVM. In Chapter 7, the MA System is introduced. Mod_AlexNet, the first contribution, an improved version of optimal deep learning model of the DE System, is compared to classic AlexNet in terms of performance across various optimizers and batch sizes. An enhanced system with feature fusion and selection integrating a majority voting ensemble classification model is detailed in Chapter 8. The system integrates feature

extraction, fusion, and selection techniques within a voting ensemble model. Despite improvements in abnormality class discrimination, further modifications were considered for enhanced performance. Chapter 9 introduces a three-layer Multi-Head Self-Attention model, which includes a novel self-attention model, MHMA model, designed to address challenges associated with variable breast size and density. Optimisation on various parameters was conducted, and results were analysed to assess the significance of the contribution to improvement.

A hybrid Multi-Head self-attention model with feature fusion, feature selection, and a novel ensemble classification model is introduced in Chapter 10, and it incorporates all previous contributions into an integrated ensemble classifier. The final contribution involves developing an ensemble classifier, IVECM, allocating weights to each class, and considering several weights from each classifier. The thesis wraps up in Chapter 11 with conclusion, including key findings and recommendations for further research.

Chapter 2 Background Medical Domain Knowledge

2.1 Introduction

A full description of the medical background will be given in this chapter. Initially, a thorough anatomy of the breast will be discussed, detailing the nuances of its structure. This will cover the description of various tissue types that comprise the breast in addition to a close examination of the characteristics of the breast tissue. Following that, a thorough explanation of the many types of breast abnormalities will be provided, along with the unique traits associated with each abnormality. In a later subsection, the many forms of breast calcifications, their signs, and how they appear on various imaging modalities will all be reviewed.

A detailed explanation of the concepts, uses, and advantages of breast tomosynthesis—a sophisticated imaging method—will be given, along with a review of the available imaging modalities for assessing the breast. This structured discussion aims to provide a comprehensive foundation for the analysis and discussions that will come in this dissertation.

2.2 Breast Structure

The mammary gland, an essential component of the female anatomy, is the tissue that covers the pectoral muscles in the chest area. Women's breasts are made up of specialised glandular tissue that produces milk and fat tissue that controls the size of the breast. In addition to providing appropriate support, ligaments and connective tissue contribute to structural integrity and form. Neural innervation provides sensory perception to the breast (Bistoni and Farhadi, 2015). In addition, the vascular network of the breast is made up of linked lymph nodes, blood arteries, and lymph vessels, as shown in Figure 2.1. This is a brief description of the unique physiological characteristics of the breast.

Different ratios of glandular and fat tissue account for a considerable deal of diversity in the size and appearance of female breasts. It is important to note that breast size is rarely constant. The breasts of women and men are pretty similar when compared anatomically. However, one distinctive feature of male breast tissue is the absence

of specific lobules, which is related to the absence of a physiological need to produce milk (Bistoni and Farhadi, 2015).

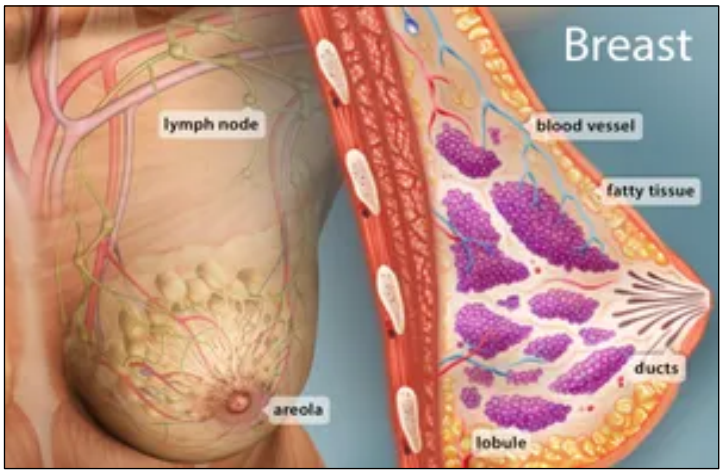

Figure 2.1 Breast Structure (WebMD, 2022)

2.3 Breast Lumps

A breast lump is a small area of internal swelling, elevation, or bulging that feels distinct from the surrounding breast tissue or from the same spot in the other breast. Breast lumps can be classified as benign (non-cancerous) or malignant (cancerous) tumours. Most breast lumps are benign and can have a variety of causes, such as cyst formation, fibroadenoma, trauma, infection, or fibrocystic disorders of the breast (Daly, C., Puckett, Y., 2020).

2.3.1 Benign Lumps

a. Hyperplasia

Hyperplasia is the abundance of cell growth that usually occurs in the breast ducts or lobules, increasing the risk of breast cancer. A growing range of screening and risk-reduction techniques are among the suggested strategies (Paepke et al., 2018).

b. Cysts

A sac filled with fluid that resembles a bump or sore area. These cysts are frequent in premenopausal women and do not raise the risk of breast cancer. Unless it causes pain, it usually doesn't need to be removed (Paepke et al., 2018).

c. Fibroadenoma

A firm, smooth, or rubbery mass within the breast tissue that is movable. The chance of developing breast cancer is not increased by this kind of lump. Most often found in females between the ages of 15 and 35, it is not removed unless it causes pain (Paepke et al., 2018).

d. Papillomas

Limited growths that may result in nipple discharges that are in the breast ducts. The presence of abnormal cells may raise the chance of breast cancer. Most women who have this benign lump removed surgically are between the ages of 35 and 55 (Paepke et al., 2018).

2.3.2 Malignant Lumps

Uncontrolled cell proliferation within the breast tissue is a hallmark of breast cancer, a malignant disorder. This disease is highly heterogeneous at the genetic level, with several subgroups exhibiting unique biological traits. It is currently the most common cancer in women worldwide. One important component of managing breast cancer concerns early-stage patients, for which curative therapies have success rates between 70% and 80% (Harbeck et al., 2019).

These tumours show in a variety of ways clinically; they might be characterised by a tender, soft, spherical mass or by an indurated, painless mass with uneven borders. The complexity of breast cancer cause is highlighted by this clinical diversity, underscoring the need for sophisticated diagnostic and treatment approaches (The American Cancer Society, 2019).

2.4 Calcification and Breast Density

2.4.1 Breast Calcification

Breast calcifications are little white spots visible on medical scans that indicate the presence of calcium deposits in the breast tissue. The presence of calcifications is a concerning sign that calls for a comprehensive assessment by medical experts. Although most breast calcifications are benign, certain irregularly shaped clusters detected on imaging scans may indicate the possibility of a malignant tumour

(Loizidou et al., 2020). Calcifications seen by breast imaging scans are divided into two groups:

a. Microcalcification

These tiny white spots are harmless and might appear in groups. When they arise in specific combinations, it may be a symptom of underlying breast cancer and require further investigation (Demetri-Lewis, Slanetz, and Eisenberg, 2012).

b. Macrocalcification

Macro-calcifications are larger, spherical, white patches that are mostly noncancerous in nature. Generally, there is no need for further testing or follow-ups when macro-calcifications are observed (Demetri-Lewis, Slanetz, and Eisenberg, 2012). Breast scans with various findings are visually represented in Figure 2.2 (National Cancer Institute, 2018).

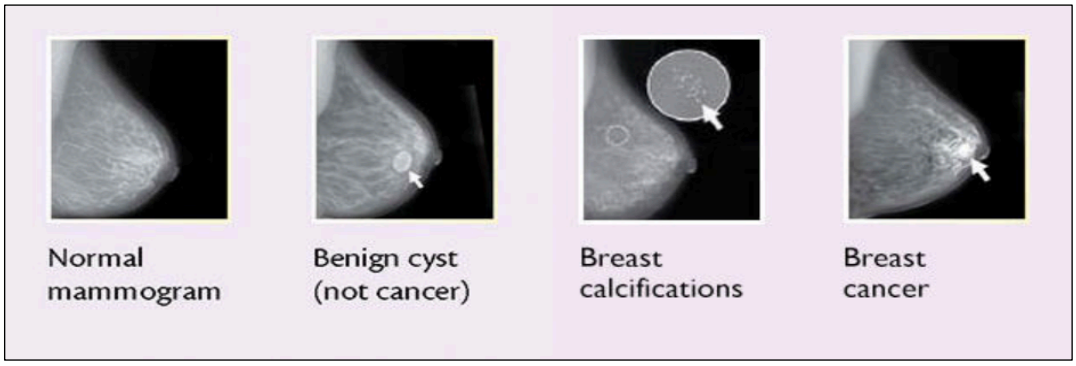

Figure 2.2 Mammogram with different findings (National Cancer Institute, 2018)

2.4.2 Breast Density

Breast density, or the thickness of the breast tissue, is a characteristic that needs to be assessed by breast scans rather than being felt. There are four classifications for breast density, ranging from primarily fatty to very dense breasts. The four categories are “almost all fatty”, “dense glandular”, “heterogeneously dense” and “extremely dense” (Saffari et al., 2020). Visual depictions of these several types of breast density are shown in Figure 2.3 (mycdiadmin, 2016).

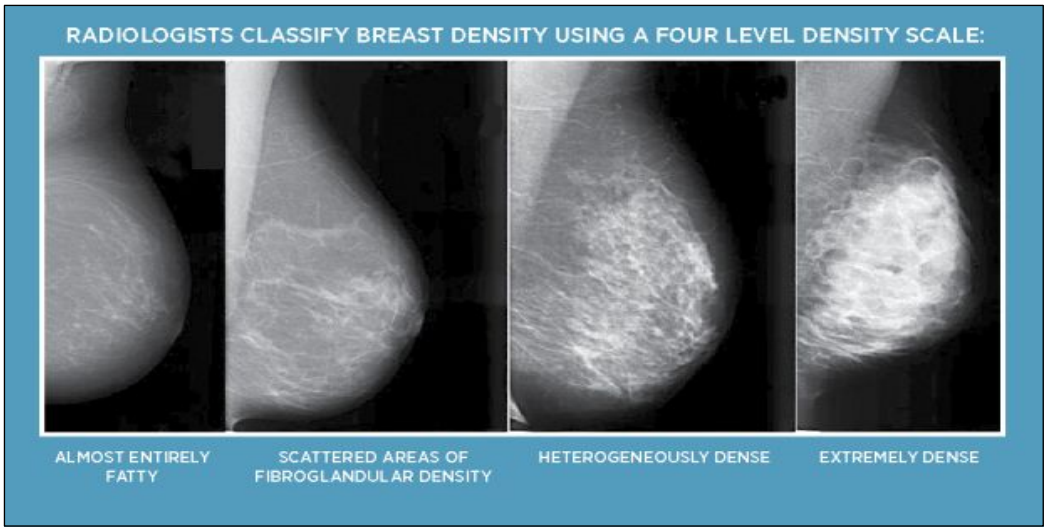

Figure 2.3 Different breast densities (mycdiadmin, 2016)

Breast density is important because it plays a part in identifying breast cancer. Those who have dense breast tissue have a higher risk of breast cancer than those whose breasts are less dense (mostly fatty). In addition, radiologists have difficulties identifying malignant abnormalities when breast tissue is dense. Dense tissues, like the one seen in Figure 2.3, show up as white on breast scans, thus any abnormalities, like possible tumours, also show up as white patches. This feature makes it more difficult to find anomalies, which emphasises how crucial it is to take breast density into account when making a diagnosis (Saffari et al., 2020).

2.5 Evaluation of Breast Lumps

A variety of modalities are utilised in breast screening with the goal of identifying breast abnormalities early on and facilitating prompt action and enhanced outcomes. Mammography is an essential tool that creates finely detailed images of the breast tissue by using low-dose X-rays. DBT, which produces three-dimensional images and provides a more comprehensive view and improves identification of minor abnormalities, enhances its proven function in screening (Iranmakani et al., 2020).

Using sound waves to provide real-time images, Ultrasound (US) is a useful addition to mammography. It is especially good at differentiating between cysts and solid masses, which helps characterise lesions that have been discovered. Magnetic resonance imaging, or MRI, is commonly used in situations that require a more thorough evaluation, such as in high-risk patients or to investigate unclear results

from other modalities, due to its ability to produce multi-dimensional images using strong magnets and radio waves (Iranmakani et al., 2020).

A new technique called elastography evaluates tissue stiffness to find possible tumours. Elastography, which measures the flexibility of breast tissue, improves diagnosis accuracy by offering more details about the type of lesions seen. However, biopsies are usually the final decision regarding the existence of a breast concern because they require the removal of tissue samples for microscopic analysis. Core needle biopsy and fine needle aspiration are typical biopsy procedures that help in accurate diagnosis and treatment recommendations, depending on the pathology of the abnormalities discovered. Combining different biopsy modalities and procedures results in a comprehensive approach to breast screening that enables more nuanced understanding of breast health and individualised patient management (Iranmakani et al., 2020).

2.5.1 MRI

Breast magnetic resonance imaging (MRI) is useful for both comprehensive visualisation and diagnosis of breast tissue. Unlike mammography and US, which rely on sound waves and X-rays, respectively, this diagnostic technique uses radio waves and magnetic fields to produce extremely detailed internal breast structure images. According to Iranmakani et al. (2020), MRI breast imaging has advantages in certain clinical circumstances since it can provide information that other imaging technologies cannot readily supply.

In addition, breast MRI is widely used in breast cancer treatment for preoperative planning. It facilitates the assessment of disease severity, tumour location and size, and surrounding tissue involvement. Using this data, surgeons may carefully design effective surgical operations, which helps patients with breast cancer receive the best possible care (Iranmakani et al., 2020).

Breast MRI is a useful diagnostic test for breast abnormalities, but because it cannot identify calcifications suggestive of breast cancer, it is rarely used as a screening procedure. Furthermore, false-positive results from MRIs often require additional testing or biopsies (Radhakrishna et al., 2018). A sample MRI scan is shown in Figure 2.4.

Figure 2.4 Example of an MRI scan (Mayo Clinic)

2.5.2 Ultrasound

Breast ultrasound is a non-invasive imaging technique that uses high-frequency sound waves to create incredibly detailed images of the breast. It is frequently employed in response to alarming results from a clinical breast exam or as a follow-up to an abnormal finding from a mammography. Sound waves are emitted by the probe of the US scanner and reflect off the tissues in the breast. A computer records the echo and uses the sound waves to create an image on the screen.

The ability of breast US to differentiate between solid and fluid-filled lumps is one of its primary benefits. This distinction can be important in the diagnosis of benign fibroadenomas and cysts. It is very helpful for ladies who have dense breasts (Hooley, Scoutt and Philpotts, 2013).

A mammography scan reveals several cancerous tumours and breast calcifications that are invisible on an US scan. Furthermore, biopsies are typically advised because an US scan is insufficient to identify whether the suspected anomaly is cancer (Hooley, Scoutt and Philpotts, 2013).

2.5.3 Elastography

A more recent development in sonographic imaging is breast elastography, which provides important information on breast diseases and anomalies. Through elasticity tests, this novel method provides vital information regarding tissue deformability and is a non-invasive way to assess the stiffness of lesions (Goddi, Bonardi, and Alessi, 2012). Even while it offers a promising path for diagnostic applications, putting it into practise has several difficulties. A significant obstacle in breast elastography is the disparity in elasticity between certain benign and malignant tumours. Due to this variance in tissue elasticity, it can be challenging to classify lesions appropriately based alone on the features of their stiffness. Another challenge is the unpredictability of transducer pressure during imaging procedures. Pressure variations can affect the quality of elastrograms and cause stiffness values of the benign and malignant tissues to overlap. According to Faruk et al. (2015), this overlap may therefore lead to increased rates of FPs and FNs during the diagnostic procedure.

Notwithstanding these obstacles, efforts are being made to improve the accuracy and overcome these limitations of breast elastography by continued study and technical improvements. It is possible to increase the usefulness of elastography in clinical settings by pursuing better imaging approaches and standardising procedures. This will give physicians more accurate information for the accurate characterisation of breast lesions.

2.5.4 Mammography

Mammography is an X-ray method that is very useful for finding breast lumps, which could be a sign of breast cancer. It is not always possible to determine with certainty if breast cancer is present or absent with a mammography scan. It enables qualified experts to determine whether the breasts have any abnormalities. To determine the most accurate diagnosis, a physical examination (looking at the breast) and further tests must be performed in addition to the mammography analysis (Hela et al., 2013).

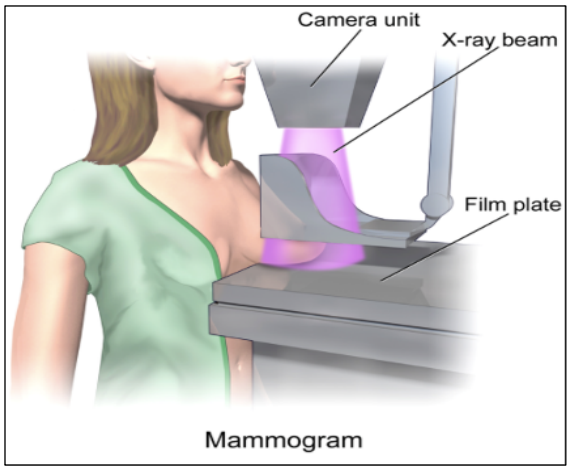

Figure 2.5 Mammogram Scan (Riggs, 2017)

Mammograms capture the differences in X-ray absorption between the various breast tissue constituents. Variations in the quantity of radiation absorbed by different tissues aid in the examination and differentiation of the details and characteristics of the tissues. Every breast is crushed on a level surface, and then an electronic device or film from an above perspective records the X-ray radiation that passes through the breast (Bandyopadhyay, 2010). The process of a mammography scan is depicted in Figure 2.5 (Riggs, 2017).

Mammograms need to have high contrast, low radiation exposure, and high resolution. High contrast mammography can be used to distinguish densities between diseased structures and normal breast tissues. To differentiate between breast calcification and soft tissues (like masses), a mammography scan needs to have high resolution. Lastly, as women typically get mammograms scanned, it is critical to have as little radiation exposure as possible during the process (IARC Working, 2016).

2.5.5 Breast Tomosynthesis

Digital Breast Tomosynthesis (DBT), also known as 3D mammography, is a modern imaging technique that makes it easier to detect breast abnormalities and cancer (Ali and Adel, 2019). Unlike the traditional 2D mammogram, which takes just one flat image of the breast, DBT captures multiple low-dose X-ray images from different angles. These images are then combined to create a detailed three-dimensional model of the breast, providing a clearer and more accurate view of the tissue (Helvie, 2010).

The process feels familiar to anyone who's had a mammogram. The breast is gently compressed and positioned, but instead of capturing one image, the X-ray tube moves in an arc over the breast, taking multiple pictures. These are processed using advanced algorithms to create a 3D model that helps doctors make more precise diagnoses (Helvie, 2010).

DBT became widely available in the U.S. on February 11, 2011, when the FDA approved the first DBT unit for clinical use, marking a major milestone in breast imaging technology (Fda.gov, 2018). One of the unique aspects of DBT is that the number of image slices taken during the scan adjusts based on the thickness of the breast, ensuring the imaging is tailored to each patient (Conant, 2014).

This breakthrough technology offers a significant step forward in breast health, giving doctors better tools to detect potential issues early. A typical DBT scan provides a much clearer picture, as illustrated in Figure 2.6 (Themes, 2016).

One significant advantage of DBT is the improved visibility of breast lesions. DBT lessens the superimposition of overlapping breast tissues by taking images in a sequence of slices, making it easier to see small abnormalities such as masses and calcifications. By allowing radiologists to examine the breast tissue in greater depth and making a more accurate distinction between benign and malignant tumours, this three-dimensional method helps to improve diagnostic accuracy.

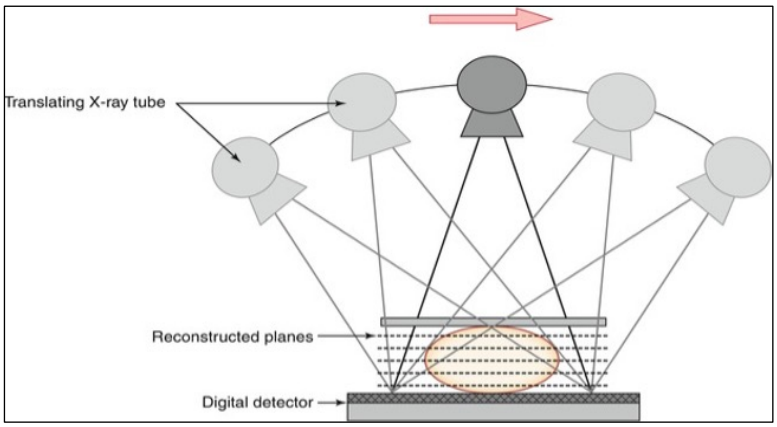

Figure 2.6 Digital Breast Tomosynthesis scan (Themes, 2016)

Compared to conventional mammography, DBT has shown promise in lowering recall rates. Layered examination of the breast reduces the need for invasive procedures or additional imaging investigations by enabling a more comprehensive evaluation of observed abnormalities. This feature lessens patient anxiety related to needless follow-up exams and helps to facilitate a more effective diagnostic process (Magni et al., 2023).

Moreover, the restricted accessibility and related expenses provide obstacles to the broad use of DBT. Some medical facilities may not have included DBT in their screening programmes, which could limit the access of the patient to the advantages of this cutting-edge imaging technique. Furthermore, radiologists must receive specialised training in interpreting DBT images because the distinct features of three-dimensional reconstructions necessitate a sophisticated comprehension for precise diagnosis (Williams and Drew, 2019).

To sum up, DBT has a lot to offer in terms of better lesion visibility and diagnostic precision for breast cancer screening. Moreover, DBT detects more breast cancer cases than traditional mammograms. As a result, it is considered a promising method to enhance both the sensitivity and specificity of mammograms (Rosenqvist, Brännmark and Dustler, 2024).

2.6 Summary

In this chapter, a comprehensive medical overview of breast anatomy was presented, outlining the composition of the breast. Next came a detailed study of breast tumours, defining their physical characteristics and encompassing a variety of forms. An examination of the various types of calcifications and the spectrum of breast densities—from very fatty to highly dense—were covered in the chapter. It was highlighted how crucial calcium deposits are to the detection of anomalies in scans and how they aid in diagnostic processes. The chapter also included a thorough examination of a variety of breast imaging modalities, outlining the advantages and disadvantages of each modality in order to diagnose breast issues. The detailed discussion focused on the complexity of detecting anomalies and supporting accurate diagnosis by paying close attention to breast density and calcifications in imaging studies.

CAD Systems is explored in the next chapter, which offers a smooth transition from conventional diagnostic methods to state-of-the-art technology applications. Illustrative cases demonstrate the vital role that CAD systems perform in the medical domain. Some instances of how these technologies help radiologists identify anomalies and work together to improve diagnosis results are described. This study highlights CAD systems as critical components in the development of intelligent medical imaging and breast diagnostic systems.

Chapter 3 AI in Medical Applications and Methodologies

3.1 Introduction

Medical applications have been transformed by CAD Systems, which are now essential diagnostic tools for medical professionals. These technologies carefully examine medical images using complex algorithms and artificial intelligence (AI) to help identify and understand possible abnormalities. When it comes to breast scans, CAD Systems have emerged as a key component in the early identification of breast cancer, significantly enhancing both patient outcomes and diagnostic accuracy (Yeasmin, 2023).

AI encompasses machine learning and deep learning, which use a variety of statistical, probabilistic, and optimisation techniques to interpret data and improve performance. These methods are quite effective, particularly when handling large and complex datasets. A subset of machine learning called deep learning models are very efficient at identifying complex patterns and extracting subtle features. These sophisticated models demonstrate their competence in medical image analysis not only by differentiating between normal and abnormal medical scans, but also by identifying minute characteristics that contribute to improved diagnostic precision. Deep learning models are important for expanding the possibilities of AI applications in the healthcare industry because of the degree of comprehension they are able to accomplish (Sarker, 2021).

The following sections offer a brief overview of the role of CAD systems in medical applications, with a specific focus on breast scans. Subsequently, the methodologies applied in the development and integration of the five systems will be presented in detail.

3.2 Computer-Aided Diagnosis Systems in Medical applications

CAD System implementation in breast scans requires an integrated approach. These devices analyse mammograms using pattern recognition algorithms to spot minute abnormalities that could be signs of breast cancer in its early stages. CAD Systems function as a powerful second opinion, helping radiologists identify lesions or microcalcifications that could be missed in a manual inspection and raising the overall sensitivity of breast cancer screenings (Chan, Samala and Hadjiiski, 2020).

The accurate diagnosis of breast cancer is enhanced by the diagnostic process that is produced by combining human expertise and technology (Yeasmin, 2023).

The ability of CAD Systems to provide statistical assessments is one of its main advantages for breast scans. These algorithms can do more than just identify possible abnormalities; they can also provide statistical measurements of lesion attributes including size, shape, and texture. Radiologists can improve their analysis and make better judgements on the type of abnormalities they find with the help of this extra layer of data (S. Arun Kumar and S. Sasikala, 2023). Furthermore, by lowering FPs and FNs, CAD Systems help to reduce the possibility of incorrect diagnosis and needless medical procedures.

CAD Systems are useful not just for initial diagnosis but also for therapy evaluation and long-term tracking. These systems demonstrate the ability to track changes in lesions over time during follow-up scans, which helps medical practitioners assess treatment outcomes and modify approaches to intervention as necessary. This long-term study makes personalised medicine more feasible by customising treatment plans to meet the demands of each patient (Petrick et al., 2013). Essentially, CAD Systems are crucial for early detection of breast cancer as well as for the continuous supervision and tracking of the condition, both of which greatly enhance the quality of patient treatment.

Based on the findings of the literature study, CAD Systems in breast scans have several applications besides their main functions of tumour classification and detection. The categorisation of breast density, a critical component affecting cancer risk and screening results, is one significant application. Healthcare practitioners can better adjust screening strategies based on individual breast structure by using CAD Systems to precisely classify breast tissue density. These devices also do a great job of detecting both micro and macro calcifications, which can reveal important information about possible breast cancer beginnings. In addition, CAD Systems aid in the classification of results according to the Breast Imaging Reporting and Data System (BI-RADS), improving reporting accuracy. This multifunctional approach expands the application of CAD Systems and demonstrates their adaptability to several aspects of breast health in addition to the main goal of tumour identification and classification.

In conclusion, CAD Systems have developed into important technologies in the field of medicine, especially for breast scanning. Their ability to improve the precision of diagnosis, offer evaluations, and enable complete patient care highlights their importance in the advancement of breast health. As technology develops, the incorporation of CAD Systems into standard clinical practice signifies an important shift towards more accurate, effective, and customised healthcare treatments concerning the diagnosis and treatment of breast cancer.

The methodologies employed in the integrated models in CAD systems are covered in the section that follows. This part offers a thorough overview and in-depth explanation of the different methodologies applied in each of the five developed systems this thesis covers. Every methodology is examined to provide an in-depth understanding of how it is applied and how it contributes to the overall efficiency of the systems. The purpose of this thorough analysis is to provide an in-depth knowledge of the specifics of the employed approaches, clarifying their purposes and importance within the framework of the thesis.

3.3 Methodologies implemented in Integrated Models in Computer-Aided Systems

A multimodal strategy encompassing multiple critical steps is essential for attaining reliable and precise results in CAD Systems for medical image classification. Augmentation is usually the first step in the workflow, where the dataset can be diversified to increase model robustness by exposing it to a wider range of changes. After the augmentation stage, the pre-processing model usually follows. In this model, the raw medical images are enhanced, normalised, and noise-reduced to prepare the data for further analysis. Subsequently, the feature extraction stage involves identifying and capturing relevant patterns and structures from the enhanced augmented data. Since deep learning models are so effective at automatically learning hierarchical features, this stage is frequently carried out using convolutional neural networks (CNNs) or DCNNs. Feature selection approaches can be utilised after feature extraction to improve model efficiency and lower dimensionality by keeping the most useful features. Finally, a classification model, which may be ranging from complex deep learning architectures to more conventional machine learning methods like SVM, is fed the features that have been

selected: pre-processing, augmentation, feature extraction, feature selection, and classification are all iterative processes that together enable CAD Systems to produce precise and complex medical image classifications, promoting better patient care and diagnostics.

3.3.1 Data Augmentation

Data augmentation is a technique for enhancing the amount and complexity of current data on purpose. Data augmentation has become a major study issue in the field of deep learning in recent years. Deep learning requires many training samples, yet available datasets in the medical field have limited resources. A data augmentation phase is necessary to increase the variety of the original dataset (Perez et al., 2017). The most well-known data augmentation techniques are:

- Flipping: creates a mirror copy of the original image.
- Rotation: the process of rotating an image around its centre pixel.
- Noise addition: adds noise to an image.

In the developed research, a series of image augmentation techniques were strategically applied to enhance the robustness of the training dataset. Flipping once vertically and once horizontally was the first augmentation technique. After the image was flipped, random brightness level adjustments were applied between -0.3 and -0.1. The augmentation technique was utilised to replicate different lighting circumstances, which is an essential component in real-life situations. The model performs better in a variety of lighting conditions as a result of the brightness reduction since it enhances the generalisation ability of the model across various brightness levels.

3.3.2 Pre-processing

Before raw medical imaging data enters the deep learning model, it undergoes a series of techniques known as pre-processing in the framework of tomosynthesis classification. A key component of pre-processing is image enhancement, which involves performing a number of enhancements to image quality, and clarity in order to improve medical imaging. These methods including noise reduction, histogram equalisation, and contrast enhancement. The main advantage of employing pre-processing in tomosynthesis classification is its ability to tackle challenges with

medical imaging data, including variations in brightness and noise. Pre-processing, thus, serves as an essential model in the enhancement of noisy tomosynthesis images, creating the framework for improved efficiency and diagnostic precision in classification models. The essential role of pre-processing is guarantying that the next phases of the deep learning process are more capable of extracting significant features from the complex tomosynthesis datasets, hence improving the accuracy of medical image classification. Different pre-processing techniques were applied to the five systems developed in the context of this thesis. Different enhancement techniques and colour mapping models were applied in every system. The techniques employed, along with their corresponding explanations, are detailed below.

3.3.2.1 Enhancement techniques

Within the area of tomosynthesis image analysis, the application of diverse image enhancement techniques is essential for optimising image quality, which in turn boosts classification model performance.

i. **Contrast Limited Adaptive Histogram Equalisation (CLAHE)**

An effective method for tomosynthesis classification, demonstrating its effectiveness through an advanced approach for image enhancement. Based on local features in the image, CLAHE dynamically adjusts the intensity distribution to dynamically address fluctuations in pixel intensity. This process is based on the adaptive histogram equalisation principle. This adaptable quality is especially helpful in reducing problems caused by irregular intensity distributions, which are a common problem in tomosynthesis images. To improve local contrast without boosting noise, CLAHE operates by segmenting the image into smaller areas, determining the histogram for each region, and then locally performing contrast enhancement (Singh, Mukundan and De Ryke, 2019). When it comes to tomosynthesis, the ability of CLAHE to adaptively improve histogram adjustment locally is essential for highlighting small and structural details, which enhances the visibility of features that are essential for precise classification. CLAHE is an important model for improving image quality and raising the precision of classification models in medical imaging due to its ability to adjust to the specific characteristics of medical images and how well it fits with the complexity of tomosynthesis data (Alshamrani et al., 2022).

ii. **Adaptive filtering**

A complex technique for feature enhancement which operates by dynamically altering filter parameters in response to local image features. This method is especially useful in the field of tomosynthesis since it can adjust to different areas of an image. Through the use of adaptive filtering for distinct areas, the enhancing procedure is carefully customised to the unique properties of different tissue configurations. This adaptability guarantees a precisely adjusted strategy, enabling the algorithm to concentrate on and highlight specific areas of interest. Especially when it comes to tomosynthesis images, where there are many small variations in tissue density and structure, adaptive filtering greatly improves the ability of the model to discriminate. The ability of the technique to distinguish tiny differences in different regions of the tomosynthesis images plays a crucial role in increasing the overall diagnostic accuracy of the classification model, making it an essential part of the image enhancement process (Kikinis and Knutsson, 2000). Gaussian and Weiner adaptive filters are applied in the context of tomosynthesis classification as a method for addressing different image enhancement challenges. The Gaussian adaptive filter plays a crucial role in efficiently reducing noise while maintaining important image details. It does this by employing a kernel with adaptable weights based on local image attributes. The Weiner adaptive filter reduces additive noise according to the local signal-to-noise ratio while simultaneously adjusting its parameters according to the local features of the image. Combining these filters results in a thorough enhancing approach for tomosynthesis images that focuses on noise reduction. This dual-filter strategy attempts to maximise image quality, consequently offering a better and more detailed representation of anatomical features and abnormalities (Kikinis and Knutsson, 2000).

3.3.2.2 Colour Mapping

Hue, Saturation, and Value (HSV) colour mapping is an approach to express colours in a colour space. The HSV colour model divides colour information into three distinct components: hue, saturation, and value. This is in contrast to the RGB colour model, which expresses colours as combinations of red, green, and blue. Moreover, it proved to be a crucial component of image processing; image enhancement and categorising are two areas in which it is very useful.

Hue is a visual characteristic of colour that has values between 0 and 360 degrees. This component is symbolically represented by the colour wheel, which assigns 0 and 360 degrees to red, 120 degrees to green, and 240 degrees to blue. The unique colour identity is defined by the assignment of particular hues within this spectrum. Saturation is a percentage that ranges from 0% (grayscale or absence of colour) to 100% (highest intensity), and it impacts the dominancy or intensity of colour. A finer control over the colour spectrum can be achieved by adjusting the saturation component, which affects the colour intensity (Fleyeh, 2008). Value is expressed as a percentage and describes the hue or intensity of a colour. The value component controls the overall luminance of the colour, with 0% indicating black (lowest brightness) and 100% indicating maximum brightness. This parameter can be varied to change the brightness characteristics (Mmed, 2023). Figure 3.1 displays the HSV colour space.

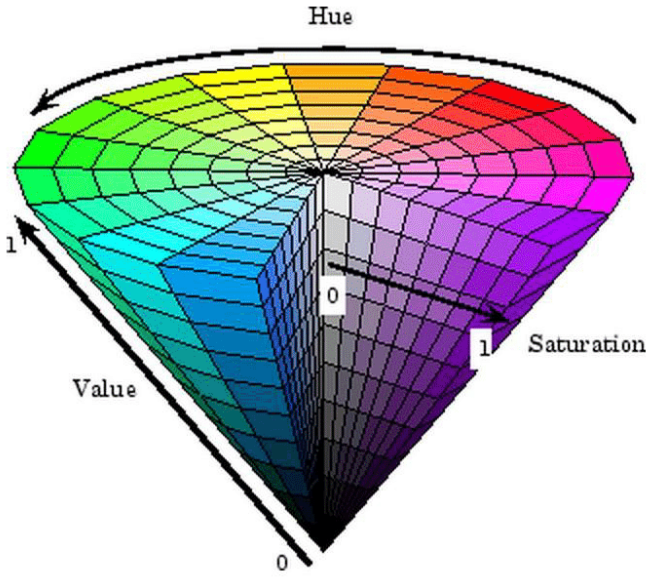

Figure 3.1 HSV colour space (Erdogan and Yilmaz, 2014)

Figure 3.1 shows the cylindrical architecture of the HSV colour space. This architecture is well-suited for simple understanding and use in a variety of image processing applications. Additionally, adjustments are made easier by separating colour information (Hue and Saturation) from intensity (Value). Because of this distinction, brightness may be independently adjusted, offering a useful tool for improving images without sacrificing colour details. This feature is especially helpful

for accurate intensity representation, including tissue separation in medical images. The accurate handling of colour information within the HSV colour space is another benefit.

HSV provides a significant advantage over the RGB model in handling lighting variations. In the RGB model, brightness and colour are interdependent, which means that adjusting brightness affects the colour balance, and changing colour intensity impacts brightness. This coupling makes it difficult to perform precise brightness and contrast adjustments without distorting colour information. In contrast, HSV separates brightness (value) from colour (hue and saturation), allowing independent adjustments to brightness without compromising colour accuracy. This ensures that hue and saturation remain stable under different lighting conditions, which guarantees more consistent colour representation and improves classification accuracy.

The RGB model is also limited in its ability to represent complex colour variations accurately in medical imaging. In tomosynthesis scans, for example, subtle differences in tissue density and brightness are critical for distinguishing between normal, benign, and malignant cases. The HSV model's separation of brightness and colour information allows for more refined control over contrast and feature extraction, enhancing the accuracy of classification. Unlike RGB, where changes in brightness can distort colour interpretation, HSV preserves the original colour integrity, ensuring that classification remains consistent even when lighting conditions vary.

This adaptability gives HSV a significant advantage in tomosynthesis-based breast cancer classification, where variations in tissue density and brightness often complicate accurate diagnosis. The ability to independently adjust brightness while preserving colour integrity makes HSV a powerful tool for improving classification accuracy and overall diagnostic performance. HSV's combination of brightness control and colour stability makes it particularly well suited for complex medical imaging tasks where colour accuracy and contrast are essential.

3.3.3 Feature Extraction

The feature extraction step is the process by which a large set of raw data is reduced to smaller groupings for simpler processing. In order to gather beneficial characteristics that differentiate between different classes and produce a more accurate classification, feature extraction is a crucial step. It is possible to extract features using both linear and non-linear techniques. Low-level features like edges, textures, and colour histograms are examples of linear features. Conventional computer vision methods like edge detection, texture analysis, and colour histogram analysis can be used to extract these features (Ghojogh et al., 2019). On the other hand, non-linear features are higher level features including shapes, objects, and patterns. More sophisticated machine learning methods, such as, CNNs can be used to extract these features. These features are more complex and usually require more time and computing resources to extract but can provide more powerful features for image classification tasks (Chen et al., 2016).

The primary difference between linear and nonlinear features is their identification and accuracy. When comparing linear versus non-linear features, linear features are typically easier to identify and provide more accurate results. This is explained by the fact that linear features are easier to understand and use because they are expressed as a weighted sum of input variables. Conversely, compared to linear features, non-linear features might provide more detailed information about an image. As a result, non-linear features are useful for identifying deeper features of an image that linear features could miss. This research involved the extraction of many linear and non-linear features, including linear features derived from HOG descriptors and non-linear features derived from deep learning models.

3.3.3.1 Deep learning models

Deep learning models have gained prominence in image classification and identification, and academics are working to enhance performance by constructing deeper learning models. Several state-of-the-art deep learning models were employed in this thesis to enhance the feature extraction process. The utilised deep learning models encompassed AlexNet, ResNet-18, GoogleNet, MobileNetV2, DenseNet-201, VGG-16, SqueezeNet, and ResNet-50. The sections that follow offer a thorough explanation of these models.

i. AlexNet

AlexNet is an 8-layer CNN built in 2012 by Alex et al. (Krizhevsky, Sutskever and Hinton, 2012), it took part in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012 and achieved a top five error rate (Krizhevsky, Sutskever and Hinton, 2012). AlexNet outperformed all standard machine learning and computer vision algorithms in terms of recognition accuracy. It was a milestone event in the history of machine learning and computer vision for visual identification and classification tasks, and it signalled the beginning of a spike in interest in deep learning (Alom et al., 2018).

The network is composed of up of eight layers: five convolutional layers, followed by max-pooling layers, and three FC layers. It is significant for using the Rectified Linear Unit (ReLU) activation function, which accelerates training convergence and assists in reducing the vanishing gradient issue. Figure 3.2 displays the architecture of AlexNet.

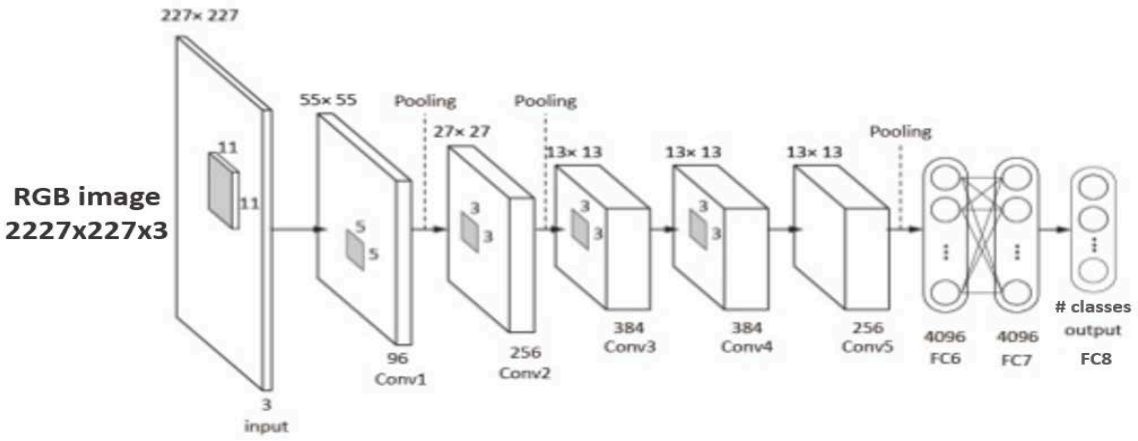

Figure 3.2 AlexNet architecture (Khvostikov et al., 2018)

To help extract high-level features, the first convolutional layer of AlexNet has a kernel size of 11x11 with a stride of 4. Subsequent convolutional layers with ReLU activations contribute to the nonlinearity of the model, enabling it to learn more complex patterns. The max-pooling layers that follow these convolutional layers down sample the spatial dimensions of the feature maps, allowing the network to efficiently collect both local and global features. A distinctive feature of AlexNet is the addition of Local Response Normalisation (LRN) after the first and second convolutional layers. LRN normalises the responses between neighbouring neurons,

which encourages the discovery of invariant features and improves the network's generalisation ability. This enhances the model's discrimination capability by increasing the sensitivity of neurons to locally important features while suppressing less informative activations, ultimately improving classification performance (Krizhevsky, Sutskever and Hinton, 2012).

The FC layers at the final stage of the network generate the hierarchical features extracted by the convolutional layers. These FC layers are subjected to dropout regularisation, which helps to avoid overfitting during training. This regularisation approach includes randomly "dropping out" neurons during training, causing the network to learn more robust features. ReLU activations, LRN, and dropout regularisation are three strategies that work together to reduce the vanishing gradient issue and speed up training convergence (Krizhevsky, Sutskever and Hinton, 2012). Being one of the first large-scale CNNs, AlexNet introduced a significant number of parameters, which contributed to its success.

AlexNet is able to capture complex hierarchical information using its deep architecture, which allows it to be effective at image classification tasks. The mix of convolutional layers, nonlinear activations, normalisation approaches, and regularisation algorithms has since served as an inspiration for succeeding CNN architectures. AlexNet left an eternal mark by laying the groundwork for the creation of more complex and deep neural networks. Moreover, it has been frequently adopted due to its simple network construction and shallow depth (Yan, Jing, and Wang, 2021).

ii. ResNets

ResNet, also known as deep residual network, is an Artificial Neural Network (ANN) model developed in 2016 by Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (He et al., 2016). The introduction of residual blocks, also known as residual units, which enable the network to learn residual functions, constitutes the main novelty of ResNet. These blocks have skip connections, sometimes referred to as identity shortcuts, which allow information to move directly from one network layer to a subsequent layer. Instead of learning the complete transformation, the objective is to learn the residual mapping, which is defined mathematically as the difference

between the input and output of a layer. By addressing the vanishing gradient issue, this method makes it easier to train deeper networks (He et al., 2016).
In ResNet-18, the residual block usually consists of two convolutional layers that are sequentially followed by ReLU activation and batch normalisation. By skipping these convolutional layers, the skip connection adds the input straight to the output of the residual block. This facilitates the convergence of the network by enabling the model to learn the residual rather than the complete transformation. The architecture and layers parameters of ResNet-18 are demonstrated in Figure 3.3 and Table 3.1.

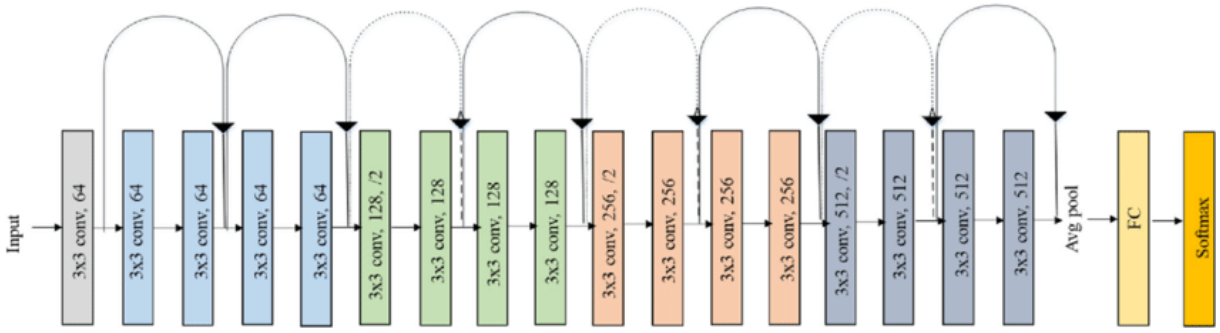

Figure 3.3 ResNet-18 architecture (Ramzan et al., 2019)

As shown in Figure 3.3, the network consists of 18 layers in total: a fully-connected layer, an additional softmax layer used for the classification task, and 17 convolutional layers. The network is constructed so that layers with the same output feature map size have an equal number of filters, and the convolutional layers use 3 × 3 filters. On the other hand, the number of filters in the layers doubles when the output feature map is cut in half. The down sampling is done using convolutional layers with a stride of 2. An average-pooling layer, a fully-connected layer, and a softmax layer for classification are located towards the final layers of the architecture.

Table 3.1 ResNet-18 architecture and layers parameters (Bui Hai Phong, Thang Manh Hoang and Le, 2020)

Layer Name	Output Size	Layer parameters
$conv1$	$112\times112\times64$	7×7,64, stride 2
$conv2\_x$	$56\times56\times64$	$3\times3 maxpool, stride 2$ $\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix} \times 2$
$conv3\_x$	$28\times28\times128$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix} \times 2$
$conv4\_x$	$14\times14\times256$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix} \times 2$
$conv5\_x$	$7\times7\times512$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix} \times 2$
average pool	$1\times1\times512$	7×7 average pool
fully connected	2	512×2 fully connections
softmax	2	Classification results

Residual shortcut connections are integrated between layers across the whole network. There are two kinds of these linkages. The first kind, shown by solid lines, is applied in cases where the size of the input and output are the same. When the dimensions rise, the second type—shown by dotted lines—is used. This connection maintains identity mapping in spite of the increased dimensions; however, it uses a stride of 2 and zero padding for the higher dimensions (Ramzan et al., 2019).

The two variations of the ResNet (Residual Network) design, ResNet-18 and ResNet-50, differ primarily in their number of layers and depths. With only 18 layers, ResNet-18 is a shallower network than ResNet-50. It is made to leverage the advantages of residual connections while offering a lighter option. Each fundamental building pieces of the architecture consists of batch normalisation, shortcut connections, and convolutional layers. On the other hand, ResNet-50 offers more representational capacity because it is a deeper network with 50 layers. ResNet-50 can extract more intricate features and hierarchies from the data at this depth. The architecture allows the network to learn complex patterns and representations by including residual blocks with several convolutional layers. ResNet-50 is frequently used in applications like image classification on huge datasets because it performs well on tasks demanding a better comprehension of visual data (Sarwinda et al., 2021).

When compared to other architectural models, the ResNet model has the distinction of retaining performance even as the design grows to be more complicated. One of the benefits of ResNet-18 is its capacity to train extremely deep networks efficiently. The learning of identity mappings is made possible by the addition of residual blocks, which speeds training and lets the network benefit from greater depth. It has been demonstrated that this architectural layout lowers the possibility of vanishing gradients, facilitates optimisation, and eventually boosts deep network performance (Sarwinda et al., 2021). The ResNet model beats other models in image classification, implying that the image characteristics were efficiently recovered by ResNet (He et al., 2016).

iii. GoogleNet

GoogleNet is a 22-layer DCNN built by Google researchers (Szegedy et al., 2014) as a version of the inception network, and the input layer of the GoogleNet architecture processes a 224X224 image. The GoogleNet architecture was designed to be a computing powerhouse that outperformed some of its predecessors or comparable networks at the time. The first convolutional layers of the complex architecture are where basic features like edges and textures are captured using 3x3 filters. However, the distinguishing innovation is found in the inception modules. (Alsharman and Jawarneh, 2020). The structure of GoogleNet is illustrated in Figure 3.4, accompanied by an elaborate description of the layer parameters provided in Table 3.2.

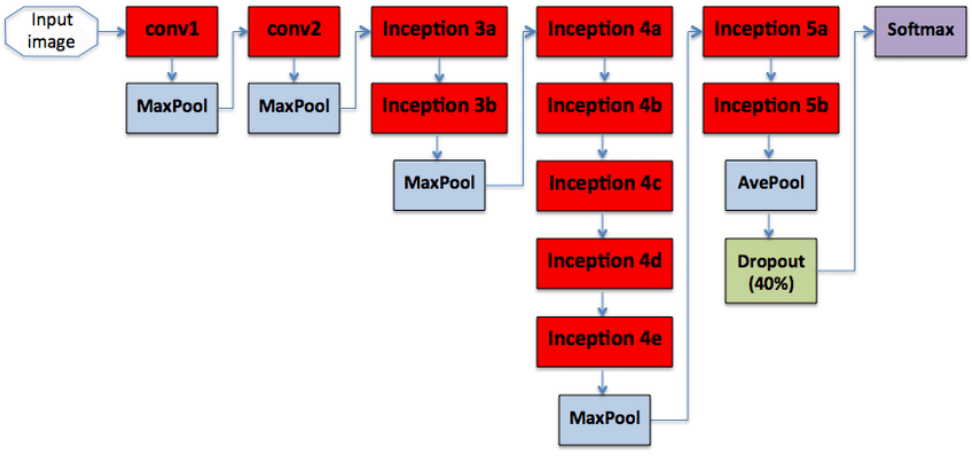

Figure 3.4 *Architecture of GoogleNet, where all convolutional layers and inception modules have a depth of two (Pawara et al., 2017)*

Basic components of GoogleNet, called inception modules, combine multiple-sized filters (1x1, 3x3, 5x5) with a max-pooling layer to transform feature extraction. By using this innovative method, the network is able to effectively capture spatial hierarchies, which promotes multi-scale feature representation and computational efficiency. The 1x1 convolutions, also known as bottleneck layers, optimise computational resources by limiting the number of input channels prior to more complex convolutions, a design choice that considerably improves the overall efficiency.

Table 3.2 GoogleNet Layer Architecture information (Szegedy et al., 2014)

type	patch size/ stride	output size	depth	#1×1	#3×3 reduce	#3×3	#5×5 reduce	#5×5	pool proj	params	ops
convolution	7×7/2	112×112×64	1							2.7K	34M
max pool	3×3/2	56×56×64	0								
convolution	3×3/1	56×56×192	2		64	192				112K	360M
max pool	3×3/2	28×28×192	0								
inception (3a)		28×28×256	2	64	96	128	16	32	32	159K	128M
inception (3b)		28×28×480	2	128	128	192	32	96	64	380K	304M
max pool	3×3/2	14×14×480	0								
inception (4a)		14×14×512	2	192	96	208	16	48	64	364K	73M
inception (4b)		14×14×512	2	160	112	224	24	64	64	437K	88M
inception (4c)		14×14×512	2	128	128	256	24	64	64	463K	100M
inception (4d)		14×14×528	2	112	144	288	32	64	64	580K	119M
inception (4e)		14×14×832	2	256	160	320	32	128	128	840K	170M
max pool	3×3/2	7×7×832	0								
inception (5a)		7×7×832	2	256	160	320	32	128	128	1072K	54M
inception (5b)		7×7×1024	2	384	192	384	48	128	128	1388K	71M
avg pool	7×7/1	1×1×1024	0								
dropout (40%)		1×1×1024	0								
linear		1×1×1000	1							1000K	1M
softmax		1×1×1000	0								

To address training challenges associated with very deep networks, GoogleNet introduces auxiliary classifiers at intermediate layers during training. By connecting auxiliary classifiers to these intermediate layers, they aimed to encourage discriminating in the lower levels of the classifier, improve the gradient signal that is propagated back, and give further regularisation. Smaller convolutional networks are placed upon the output of the Inception (4a) and (4d) modules to create these classifiers (Szegedy et al., 2014). Their loss is added, during training, to the overall loss of the network with a discount weight. These auxiliary networks are dropped at inference time. By offering more oversight, these classifiers facilitate gradient flow and minimise the impact of the problem of vanishing gradients. This thoughtful inclusion strengthens the training process, allowing for more effective acquisition of hierarchical features (Pawara et al., 2017).

Finally, a softmax activation for classification and FC layers finalise the design. The ability of the network to accurately classify images is aided by both the auxiliary classifiers and the main classifier. In addition to being exceptionally accurate, architecture of GoogleNet takes into account the processing requirements of deep networks, increasing its viability for practical implementation. One of the strengths of GoogleNet is that, in comparison to other deep architectures, it requires a lot less parameters to attain comparable performance on image classification tasks. The multi-scale feature extraction of the inception modules demonstrates the ability of the model to effectively capture complex patterns (Szegedy et al., 2014).

iv. MobileNet V2

MobileNetV2 is an important CNN architecture that is optimised for efficient image classification. Developed by Google researchers under the direction of Menglong Zhu, Andrew Howard, and Mark Sandler (Sandler et al., 2018), MobileNetV2 offers a number of enhancements to maximise model accuracy, size, and computing performance. Inverted residual blocks, a key component of the architecture of MobileNetV2, are essential to striking a balance between model complexity and efficiency. Each inverted residual block has three major components: a lightweight depth wise separable convolution, a linear bottleneck, and a shortcut link. By factorising the convolution operation into depth wise and pointwise convolutions, the depth wise separable convolution lowers the computational cost and parameter count.

As shown in Figure 3.5, the initial layer of MobileNetV2, known as depth wise convolution, performs lightweight filtering by applying a single convolutional filter to each input channel. The second layer consists of a 1 × 1 convolution, also known as a pointwise convolution, which computes linear combinations of the input channels to produce new features. The backbone of the network is made up of several inverted residual blocks that come after the first layer. These building blocks are in charge of extracting hierarchical features at various scales, which makes it easier to represent complex patterns in the input data. The inverted residual blocks include linear bottlenecks, which allow the network to capture and propagate critical information more efficiently (Sandler et al., 2018).

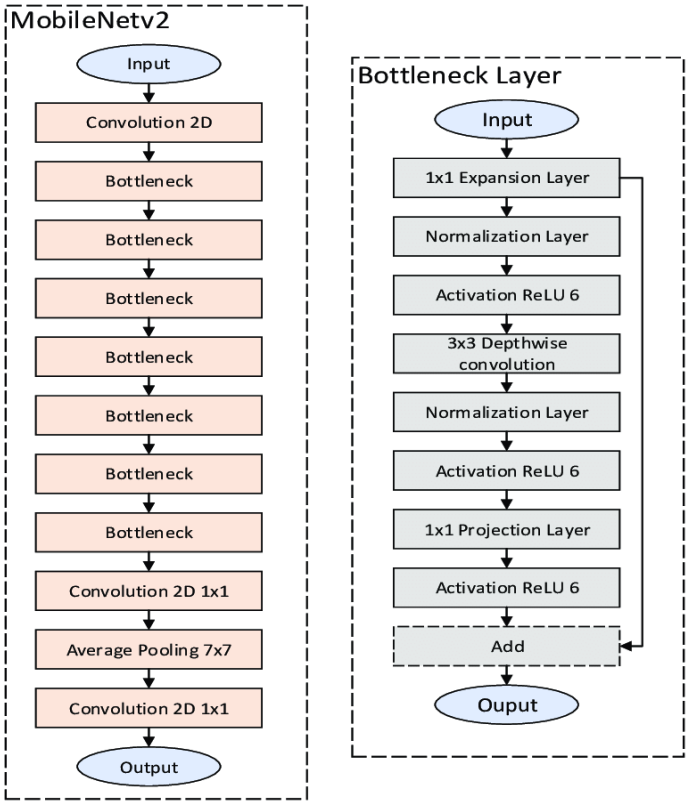

Figure 3.5 MobileNet V2 Architecture (Antonios Tragoudaras et al., 2022)

In order to promote faster convergence during training and computing efficiency, the linear bottleneck decreases the dimensionality of the input before applying non-linear activation. Through the use of a linear bottleneck and a 1x1 convolutional layer to compress the input channels, this dimensionality reduction is accomplished. To restore the feature space to its initial size, a depth-wise separable convolution and a further 1x1 convolution are performed after this. The bottleneck layer has two purposes; in the first place, it drastically lowers the computing cost of the depth wise separable convolution and other related processes. While maintaining the necessary information for precise classification, the computations become more efficient due to the decrease in the number of input channels. Second, by first compressing and then expanding the feature space, the bottleneck layer aids in preserving a balance between computational efficiency and the ability of the network to capture complicated information. Understanding parameters of each layer in MobileNetV2 is essential to figuring out how effective it is. The behaviour of each depth wise separable convolutions of inverted residual block, pointwise convolutions, and linear bottlenecks is controlled by particular parameters. The filter size, stride, input and

output channels, and the linear expansion factor of bottleneck are some of these characteristics (Sandler et al., 2018).

Compared to standard architectures, MobileNetV2 can achieve competitive accuracy on image classification tasks with substantially less parameters. Because of this, it is especially appropriate for systems with limited resources. The emphasis on linear bottlenecks and depth wise separable convolutions in the architecture makes it suitable for real-time applications and computationally efficient.

v. DenseNet-201

Modern CNN architecture DenseNet-201 is distinguished by its densely connected architecture, which encourages feature reuse and lessens the effects of the vanishing gradient issue. DenseNet-201, created by Gao Huang, Zhuang Liu, and Laurens van der Maaten (Huang, Liu and Weinberger, Kilian Q, 2016), expands on the features of DenseNet to provide better feature representation by adding additional layers.

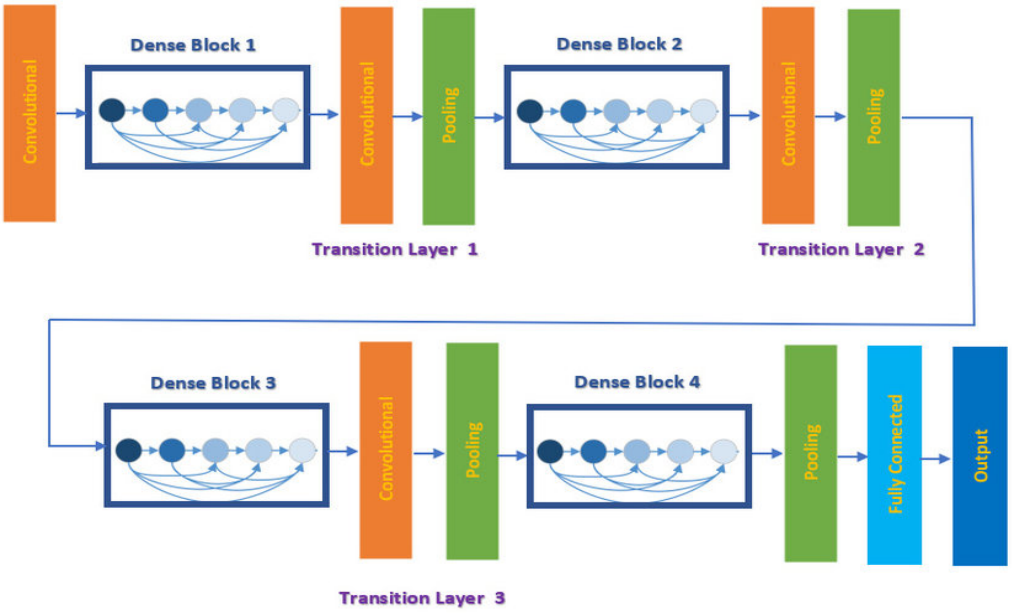

Figure 3.6 DenseNet-201 Architecture (Attallah, 2021)

Design of DenseNet-201 is distinguished by its dense blocks, as shown in Figure 3.6, each of which contains densely connected layers. Every layer in a dense block gets information from every layer that came before it, resulting in a high density of interlayer connections. Because of this design choice, the network is better able to

capture complex patterns and hierarchies by encouraging feature reuse. In order to regulate the expansion of feature maps and lower computational complexity, transition layers—which include batch normalisation, 1x1 convolutions, and average pooling—connect consecutive dense blocks (Huang, Liu and Weinberger, Kilian Q, 2016).

DenseNet-201 is composed of four dense blocks and transition layers, resulting in a densely layered network. The capacity of the model to capture complex features is influenced by the growth rate parameter, which determines how many feature maps are added in each layer inside the dense block. In addition to improving feature propagation, the densely linked architecture resolves the vanishing gradient problem, which makes training deeper networks easier (Huang, Liu and Weinberger, Kilian Q, 2016).

Table 3.3 illustrates the differences in the designs of the DenseNet models, which include DenseNet-121, DenseNet-169, DenseNet-201, and DenseNet-264. Although they utilise inter-layer connections for feature reuse and have a basic densely connected structure, their differences are primarily in the number of layers, growth rates, and total depths of the networks. For example, DenseNet-201 has more layers than DenseNet-121, which results in a higher level of model complexity and possibly stronger representational capability.

Table 3.3 The sizes of the different layers in the DenseNet DCNN (Huang, Liu and Weinberger, Kilian Q, 2016)

Layers	Output Size	DenseNet-121	DenseNet-169	DenseNet-201	DenseNet-264
Convolution	112 × 112	7 × 7 conv, stride 2			
Pooling	56 × 56	3 × 3 max pool, stride 2			
Dense Block (1)	56 × 56	[1 × 1 conv; 3 × 3 conv] × 6	[1 × 1 conv; 3 × 3 conv] × 6	[1 × 1 conv; 3 × 3 conv] × 6	[1 × 1 conv; 3 × 3 conv] × 6
Transition Layer (1)	56 × 56	1 × 1 conv			
	28 × 28	2 × 2 average pool, stride 2			
Dense Block (2)	28 × 28	[1 × 1 conv; 3 × 3 conv] × 12	[1 × 1 conv; 3 × 3 conv] × 12	[1 × 1 conv; 3 × 3 conv] × 12	[1 × 1 conv; 3 × 3 conv] × 12
Transition Layer (2)	28 × 28	1 × 1 conv			
	14 × 14	2 × 2 average pool, stride 2			
Dense Block (3)	14 × 14	[1 × 1 conv; 3 × 3 conv] × 24	[1 × 1 conv; 3 × 3 conv] × 32	[1 × 1 conv; 3 × 3 conv] × 48	[1 × 1 conv; 3 × 3 conv] × 64
Transition Layer (3)	14 × 14	1 × 1 conv			
	7 × 7	2 × 2 average pool, stride 2			
Dense Block (4)	7 × 7	[1 × 1 conv; 3 × 3 conv] × 16	[1 × 1 conv; 3 × 3 conv] × 32	[1 × 1 conv; 3 × 3 conv] × 32	[1 × 1 conv; 3 × 3 conv] × 48
Classification Layer	1 × 1	7 × 7 global average pool			
		1000D fully-connected, softmax			

Layer parameters of DenseNet-201 include details like the quantity of filters in the convolutional layers, the growth rate that determines how many feature maps are added in each dense block layer, and the compression factor that determines how many output feature maps are added in the transition layers (Attallah, 2021). The complicated parameter combinations and connectivity patterns work together to enable DenseNet-201 to capture hierarchical features with efficiency.

Compared to other DCNN, DenseNet-201 offers enhanced accuracy with fewer parameters since it maximises parameter efficiency and feature reuse. The network can better represent data because of its densely connected structure, which makes it possible to take advantage of shared information between layers. DenseNet-201 excels at image classification tasks, particularly on large datasets, where its dense connections contribute significantly to robust feature learning.

vi. VGGs

Karen Simonyan and Andrew Zisserman from the Visual Geometry Group Lab at Oxford University (Simonyan and Zisserman, 2015) suggested VGG-16 and VGG-19 in 2014. VGG-19 is a 19-layer model with many weight parameters, and VGG-16 is a 16-layer model. Figure 3.7 displays the architecture of VGG-16 DCNN and the parameters of each layer is shown in Table 3.4.

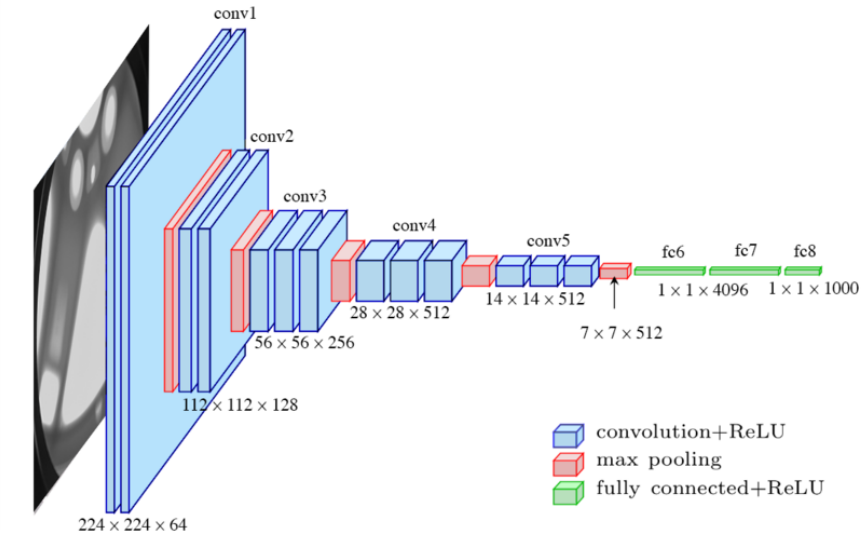

Figure 3.7 Traditional VGG-16 Architecture (Ferguson et al., 2017)

As shown in Figure 3.7 and Table 3.4, Max-pooling layers are arranged after convolutional layer blocks in the VGG architecture. Small 3x3 convolutional filters are used throughout the network to preserve a consistent receptive field, which is a crucial feature. VGG can be easily implemented and is highly interpretable due to its homogeneous structure and simplicity. Every VGG block has a max-pooling layer after a number of convolutional layers. The network can learn hierarchical features of increasing complexity thanks to this repeating structure. The benefit is that the receptive fields and filters are uniform, which encourages feature learning in an organised way (Simonyan and Zisserman, 2015). The enormous number of parameters of the VGG architecture, however, makes it frequently computationally costly.

Variants of the original VGG architecture with 16 and 19 layers, respectively, are called VGG-16 and VGG-19. These variations differ in terms of network depth but have the same block structure. With additional parameters, deeper architectures like VGG-19 are able to learn features in a more expressive way, although at the expense of greater processing power. On the other hand, VGG-16 achieves a balance between efficiency and complexity.

Table 3.4 VGG-16 network layers parameters (Zakir Ullah et al., 2021)

Layer Name	Input Shape	Output Shape	Stride	Conv Kernel Size
Conv1-1-64	$224 \times 224 \times 3$	$224 \times 224 \times 64$	1	3×3
Conv1-1-64	$224 \times 224 \times 64$	$224 \times 224 \times 64$	1	3×3
Maxpool-1	$224 \times 224 \times 64$	$112 \times 112 \times 64$	2	2×2
Conv2-1-128	$112 \times 112 \times 64$	$112 \times 112 \times 128$	1	3×3
Conv2-1-128	$112 \times 112 \times 128$	$112 \times 112 \times 128$	1	3×3
Maxpool-2	$112 \times 112 \times 128$	$56 \times 56 \times 128$	2	2×2
Conv3-1-256	$56 \times 56 \times 128$	$56 \times 56 \times 256$	1	3×3
Conv3-1-256	$56 \times 56 \times 256$	$56 \times 56 \times 256$	1	3×3
Maxpool-3	$56 \times 56 \times 256$	$28 \times 28 \times 256$	2	2×2
Conv4-1-512	$28 \times 28 \times 256$	$28 \times 28 \times 512$	1	3×3
Conv4-1-512	$28 \times 28 \times 512$	$28 \times 28 \times 512$	1	3×3
Maxpool-4	$28 \times 28 \times 512$	$14 \times 14 \times 512$	2	2×2
Conv5-1-512	$14 \times 14 \times 512$	$14 \times 14 \times 512$	1	3×3
Conv5-1-512	$14 \times 14 \times 512$	$14 \times 14 \times 512$	1	3×3
Maxpool-4	$14 \times 14 \times 512$	$7 \times 7 \times 512$	2	2×2
Fully connected-1	$1 \times 1 \times 25{,}088$	$1 \times 1 \times 4096$	1	1×1
Fully connected-2	$1 \times 1 \times 4096$	$1 \times 1 \times 4096$	1	1×1
Fully connected-3	$1 \times 1 \times 4096$	$1 \times 1 \times 1000$	1	1×1

The convolutional layers of the network, which consistently apply tiny 3x3 filters, aid in the methodical learning of features. The ability of the network to recognise complex structures is largely determined by the settings of each layer. Simplicity,

interpretability, and efficiency of VGG in capturing complex features are among its benefits. The primary disadvantage, nevertheless, is the computational expense incurred by its high number of parameters. The ability of VGG to learn hierarchical representations has allowed it to demonstrate success in image classification.

vii. SqueezeNet

Researchers at DeepScale and Stanford University (Iandola et al., 2016) have created a unique CNN architecture called SqueezeNet, which offers novel techniques for model compression without sacrificing performance. SqueezeNet is built around the concept of "fire modules," which feature squeeze and expand layers that use 1x1 convolutions to minimise parameters while maintaining expressive capability.

The first step in the fire module is the squeeze layer, which uses 1x1 convolutions to efficiently reduce the number of input channels. The parameter count of the model is reduced in a significant way because to this compression step. After the squeeze layer, the expand layer increases representational capacity by combining 1x1 and 3x3 convolutions (Iandola et al., 2016). The Fire Module is presented in Figure 3.8, demonstrating the squeeze and expand layers.

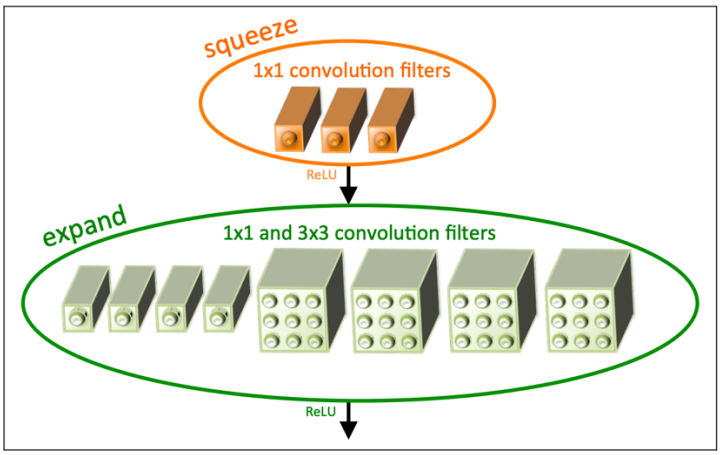

Figure 3.8 Fire Module Architecture (Iandola et al., 2016)

The 1x1 convolutions in the expand layer assist to restore expressive depth to the feature maps, whereas the 3x3 convolutions capture spatial hierarchies. By integrating these layers, the network may capture a variety of patterns in a small parameter space by making it easier to extract complex features. During the training

phase, skip connections in the fire module facilitate effective gradient flow and feature propagation in both forward and backward passes (Iandola et al., 2016).

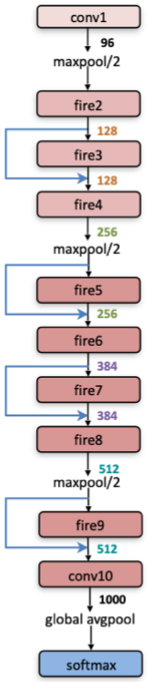

Figure 3.9 SqueezeNet Architecture (Pothos et al., 2016)

The number of filters in the squeeze and expand layers, among other fire module parameters, is an important consideration in determining the expressive power and efficiency of the model. The squeeze layer parameter known as the compression factor controls how many input channels are reduced. A careful balance is struck when configuring these parameters to guarantee that the fire module performs efficient compression without sacrificing the variety of features necessary for accurate prediction. The structure of SqueezeNet, along with the layer parameters, is illustrated in Figure 3.9, and detailed information is presented in Table 3.5.

Multiple fire modules are organised in a sequential manner within the larger structure of SqueezeNet, with max-pooling layers inserted for spatial down sampling in between. It is possible to extract hierarchical features systematically across abstraction layers because to this modular framework. With its strategic 1x1 convolutions, the fire module combination greatly enhances the performance of SqueezeNet in image classification tasks, allowing for higher accuracy at lower model sizes.

Table 3.5 SqueezeNet layers parameters (Iandola et al., 2016)

layer name/type	output size	filter size / stride (if not a fire layer)	depth	s_{1x1} (#1x1 squeeze)	e_{1x1} (#1x1 expand)	e_{3x3} (#3x3 expand)	s_{1x1} sparsity	e_{1x1} sparsity	e_{3x3} sparsity	# bits	#parameter before pruning	#parameter after pruning
input image	224x224x3										-	-
conv1	111x111x96	7x7/2 (x96)	1				100% (7x7)			6bit	14,208	14,208
maxpool1	55x55x96	3x3/2	0									
fire2	55x55x128		2	16	64	64	100%	100%	**33%**	6bit	11,920	5,746
fire3	55x55x128		2	16	64	64	100%	100%	**33%**	6bit	12,432	6,258
fire4	55x55x256		2	32	128	128	100%	100%	**33%**	6bit	45,344	20,646
maxpool4	27x27x256	3x3/2	0									
fire5	27x27x256		2	32	128	128	100%	100%	**33%**	6bit	49,440	24,742
fire6	27x27x384		2	48	192	192	100%	**50%**	**33%**	6bit	104,880	44,700
fire7	27x27x384		2	48	192	192	**50%**	100%	**33%**	6bit	111,024	46,236
fire8	27x27x512		2	64	256	256	100%	**50%**	**33%**	6bit	188,992	77,581
maxpool8	13x12x512	3x3/2	0									
fire9	13x13x512		2	64	256	256	**50%**	100%	**30%**	6bit	197,184	77,581
conv10	13x13x1000	1x1/1 (x1000)	1				**20%** (3x3)			6bit	513,000	103,400
avgpool10	1x1x1000	13x13/1	0									
	activations		parameters				compression info				1,248,424 (total)	**421,098** (total)

CNNs, like SqueezeNet, have several noteworthy benefits. Its main goal is to compress the model using 1x1 convolutions, which lowers parameters and makes the design more memory efficient. This benefit is essential for systems with limited resources. SqueezeNet strikes a balance between model simplicity and predictive performance, displaying competitive accuracy in image classification tasks despite its efficiency.

3.3.3.2 HOG

A common feature descriptor in computer vision for image classification is the HOG. By capturing local intensity gradients and their orientations in an image, the HOG approach seeks to produce a reliable representation of the textures and forms of objects. Since its first proposal in 2005 by Navneet Dalal and Bill Triggs (Dalal et al., 2005), its efficacy has led to its widespread adoption in a variety of applications, including the classification of tomosynthesis images.

There are several crucial phases involved in the extraction of HOG features. The input image is first split up into tiny, overlapping cells. Gradient information is generated within each cell by applying convolutional operators in both the horizontal and vertical directions, such as Sobel filters. Next, for every pixel in the cell, the magnitude of the gradient and orientation are determined. These gradient values are then combined to create gradient orientation histograms, where the bins in the

histogram correspond to various orientation ranges. The final feature vector is created by concatenating the histograms from each cell (Dalal et al., 2005).

Capturing local shape and texture information while being reasonably invariant to variations in light and contrast is one of the primary features of HOG features. HOG may concentrate on edges and boundaries, which are essential for differentiating between various objects or structures in an image, due to the utilisation of gradient information. The orientation histograms provide an accurate representation of the dominating gradient directions, enabling the discovery of patterns regardless of where they are located in the image (Dalal et al., 2005).

Compared to more complicated descriptors, HOG descriptors are less prone to overfitting due to their simplicity and computational efficiency. HOG can identify elements with varying sizes and orientations within an image since it is resistant to scale and rotation changes. Using local histograms enhances the discriminative power of HOG by enabling it to detect crucial features in particular areas of an image (Fleyeh and Roch, 2013).

Regarding the classification of tomosynthesis images, HOG characteristics are essential for obtaining relevant information from the 3D reconstructed slices. HOG features can be extracted from each slice to describe the distinct patterns connected to particular abnormalities or structures. Because HOG is resistant to changes in depth and orientation, it is especially useful for tomosynthesis to reliably capture texture and shape information across several slices. The method takes advantage of the fact that structures or objects of interest frequently display unique patterns in the orientations and intensities of their gradients. HOG offers a clear and informative feature vector that may be utilised for classifier training or for immediately comparing and recognising patterns in previously unknown data by expressing these patterns as histograms of gradients (Vo et al., 2013). HOG features help with the precise and effective identification of anomalies in the tomosynthesis scans.

3.3.4 Feature Fusion

Improving the overall performance of a system through feature fusion involves integrating several features from various sources or representations. For feature

fusion, there are numerous accessible methods ,each with a unique strategy and benefits.

Early fusion is a popular technique in which features from many modalities or sources are integrated at the input level prior to any additional processing. For instance, in multimodal image classification, a single feature vector is created by concatenating features from several image modalities. Early fusion provides the model with a uniform representation, which makes subsequent processing easier. That might, however, result in a higher-dimensional feature space and possibly include redundant or unnecessary data (Yao et al., 2021).

Another common strategy is called "late fusion," which combines features at a higher level of abstraction, usually after individual processing steps. Late fusion is the process of fusing features that have been independently retrieved from various sources or representations using techniques like weighted summation, concatenation, or averaging. More flexibility is possible with this method, which may also gather supporting data from many sources. On the other hand, late fusion could be sensitive to the quality of individual features and might need more complex processing for combining features correctly (Maciej Pawłowski, Wróblewska and Sylwia Sysko–Romańczuk, 2023).

Intermediate fusion is a different kind of feature fusion that involves combining features at different phases of processing. Using this method, features that were extracted at various layers of a deep neural network or processing pipeline are combined. By utilising the hierarchical structure of features acquired by deep models, intermediate fusion facilitates the combination of both high-level and low-level representations. This method can produce more resilient and discriminative feature representations by capturing both semantic information and fine-grained features (Lu et al., 2022).

This thesis implements the late fusion feature fusion, and this is achieved by concatenating features that are generated from the HOG descriptors and deep learning models, thereby combining their respective advantages. In order to capture both local texture and form attributes from HOG descriptors and high-level semantic details from deep learning models, the feature vectors obtained from each source

are combined into a single feature vector during the concatenation step. Through the integration of these features, the applied technique improves the adaptability and discriminative capability of feature representation, which improves performance in the classification of the tomosynthesis scans. The thesis improves by demonstrating the effectiveness of late feature fusion for generating richer and more complete feature representations using this fusion mechanism.

3.3.5 Feature Selection

A feature selection strategy is a vital technique used that seeks to pick out and keep the most informative and relevant features, while removing any redundant or unnecessary ones. This procedure is essential for improving understanding, cutting down on computational complexity, and optimising model performance. After feature fusion—the process of combining features from various sources to create a single, unified representation—feature selection is essential for enhancing this combined feature set, making sure that only the most informative and discriminative features are kept for further examination or training of the model for better discrimination of different classes (Tang et al., 2014).

Feature selection enhances the efficiency by choosing a collection of features that together capture the crucial attributes of the input. Furthermore, it improves the interpretability of models by concentrating on the most significant features for each class, which makes it easier to recognise the underlying correlations and patterns in the data.

The following section will examine a number of feature selection methods. These methods include chi-square, statistical tests, and minimum Redundancy Maximum Relevance (mRMR). These techniques are essential for improving the performance and fine-tuning the feature space after feature fusion.

3.3.5.1 mRMR

A powerful approach used for selecting a subset of important and non-redundant features is the Minimum Redundancy Maximum Relevance feature selection approach. Maximising the relevance of selected features to the target class while minimising their redundancy is how mRMR functions. By eliminating information duplication within the feature set, this dual goal guarantees that the features that are

selected capture the most important data (Hanchuan Peng, Fuhui Long and Ding, 2005).
Relevance and redundancy are the two primary factors that the mRMR approach determines for each feature. A measure of relationship between each attribute and the target class, such as mutual information or the correlation coefficient, is commonly used to quantify relevance. A feature that has a higher relevance score has more information about the target class in it. Redundancy, on the other hand, calculates the correlation or similarity between two sets of features. When combined, features with a high degree of redundancy may contain overlapping information and provide less information (Hanchuan Peng, Fuhui Long and Ding, 2005).

mRMR uses a heuristic method that iteratively selects features based on their relevance and redundancy scores in order to determine the optimal subset of features. The algorithm chooses the feature with the highest relevance score at each iteration, making sure the chosen feature is as minimally redundant as possible with the features that have already been selected. This method keeps on until either a predetermined stopping criterion is satisfied, or the required number of features is chosen (Hanchuan Peng, Fuhui Long and Ding, 2005).

When it comes to multi-class image classification, the mRMR approach has many benefits. First off, mRMR reduces overfitting and the curse of dimensionality by selecting features that are both relevant to the target classes and non-redundant with one another. This is especially helpful in high-dimensional feature spaces that are typical of image data. Second, by making it easier to find discriminative features that capture the distinctive qualities of each class, mRMR enhances the robustness and accuracy of the classification model (Hanchuan Peng, Fuhui Long and Ding, 2005).

3.3.5.2 Chi-square test

To find the most relevant features for classification tasks, researchers frequently employ the Chi-square feature selection technique. Usually applied to categorical data, it functions by assessing the degree of independence between each feature and the target class. Evaluating whether the observed frequency distribution from the predicted distribution under the null hypothesis of independence between the feature and target variable, is the basic idea behind the Chi-square test (Mindrila et al., 2013).

For every feature, the Chi-square statistic (also known as χ^2) is computed using the contingency table that is created from the frequencies of feature-value pairings and class labels and is shown in equation 1 (Li et al., 2017). Assuming independence, the contingency table computes predicted frequencies by cross-tabulating the occurrences of each feature value with the appropriate class labels. Next, after normalising by the predicted frequencies, the Chi-square statistic is calculated as the sum of squared differences between the observed and expected frequencies (Riffenburgh, 2006).

The mathematical formula for calculating the Chi-square statistic for a given feature is:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad (3.1)$$

Where:

- O_{ij} indicates the feature value i observed frequency of I in class j
- E_{ij} represents the predicted frequency of feature i value in class j, as determined by the independence assumption.
- The summation is performed over all feature values and class labels.

Under the null hypothesis of independence, the Chi-square statistic has a Chi-square distribution, and the number of feature values and class labels determines the degrees of freedom. A feature may be more informative for classification if there is a larger correlation between it and the target variable, as shown by a higher Chi-square value (Li et al., 2017).

The Chi-square feature selection method has several benefits when used in multi-class image classification. First, it offers a methodical and statistically valid way to assess the relevance of a feature through an examination at how it relates to a class label. Second, Chi-square feature selection enables the identification of discriminative features across several classes. Furthermore, the computational efficiency of the Chi-square test makes it adaptable to high-dimensional feature spaces, allowing for effective feature selection with minimal computational overhead. In multi-class image classification applications, the Chi-square feature selection

technique is a useful tool for improving the discriminative power and interpretability of classification models (Sumaiya Thaseen and Aswani Kumar, 2017).

3.3.5.3 F-test

The Analysis of Variance (ANOVA) F-test, usually referred to as the F-test feature selection methodology, is a statistical approach that is used to determine which features are most relevant for classification tasks. In order to choose features that show a significant difference in mean values between the classes, this technique evaluates the relevance of the variance in feature values across distinct classes (Omer Fadl Elssied, Ibrahim and Hamza Osman, 2014).

The F-test evaluates the F-statistic, which is the ratio of the variation between class means to the variance within each class. The F-statistic measures the proportion of feature value variation that can be assigned to class differences as opposed to random fluctuation within classes. Greater discriminating across classes and, consequently, a stronger relationship between the feature and the target class are indicated by higher F-statistics (Pathan et al., 2022).

Mathematically, the F-statistic (Nasiri and Alavi, 2022) for a given feature is calculated as follows:

$$F = \frac{MSB}{MSW} \qquad (3.2)$$

Where:

- *MSB* represents the mean square between classes, computed as the variance of the feature values across classes weighted by the number of samples in each class.
- *MSW* denotes the mean square within classes, calculated as the average variance of the feature values within each class.

Under the null hypothesis, which states that there is no significant variation in the mean values of the feature across classes, the F-statistic follows an F-distribution. The total number of samples and the number of classes determine the degrees of freedom for the F-distribution (Nasiri and Alavi, 2022).

The F-test feature selection method has many benefits when used in multi-class image classification. First of all, it offers a statistically valid method for assessing the ability of a feature to discriminate between classes based on those variations. Second, the F-test may be used to identify features with significant differences in mean values across many classes, making it useful. The F-test is robust to outliers and non-normality in the data. The F-test feature selection technique is a useful tool for improving the discriminatory power and interpretability of classification models.

3.3.6 Classification Model

The classification stage is a crucial phase in many machine learning and data analysis workflows. Its main job is to categorise or label incoming data according to the features that are extracted from the images. This phase usually follows the feature selection procedure, in which informative features are selected to suitably represent the data. In order to predict the class or label associated with the input, several algorithms analyse and process the input feature vector, which consists of the selected features. These algorithms can be more complex deep learning models like CNNs, Recurrent Neural Networks (RNNs), or Transformer-based architectures, or they can be more conventional machine learning classifiers like SVM, Naive Bayes (NB), K-Nearest Neighbours (KNN), and Decision Trees (DT).

Each classifier has unique strengths and weaknesses, making it appropriate for a variety of data types and classification tasks. For instance, SVMs work well in high-dimensional situations and are especially helpful when handling complex decision boundaries. Large datasets can benefit from the simplicity and speed of NB classifiers. KNN classifiers perform well with non-linear data and rely on neighbour closeness for classification. DT offer models that are easy to understand and are resistant to abnormalities and redundant features. On the other hand, deep learning models have become increasingly common due to their capacity to automatically learn hierarchical data representations; with sufficient data, these models may be able to achieve higher accuracy for more complicated tasks. In the subsequent sections, an outline of the classifiers utilised in the research conducted within this thesis across all systems is presented.

3.3.6.1 SVM

The Support Vector Machine classifier is a sophisticated machine learning algorithm that is widely used for classification problems, such as multi-class image classification. In order to divide the feature space into discrete classes and maximise the margin between them, as shown in Figure 3.10, SVM finds the best hyperplane. A selection of training samples called support vectors—which are closest to the decision boundary—determine this hyperplane (Foody et al., 2004).

The primary goal of SVM is to locate the decision boundary that maximises the difference between the support vectors of various classes. The margin denotes the separation between the nearest training samples from each class and the decision border. By maximising this margin, SVM creates a strong classification model that generalises well to new data (Foody et al., 2004).

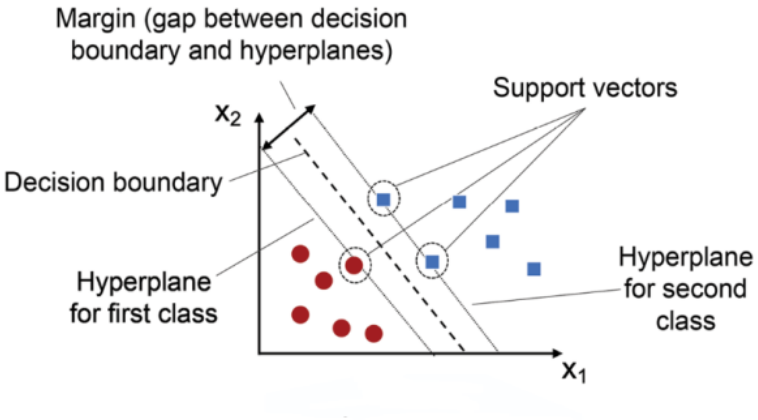

Figure 3.10 SVM Model (Kumar, 2022)

The kernel function, regularisation parameter (C), and perhaps additional kernel-specific parameters, like gamma for Radial Basis Function (RBF) kernels, are the parameters of an SVM classifier. The mapping of input features into a higher-dimensional space, where a linear decision boundary can be located, is defined by the kernel function. Sigmoid, polynomial, linear, and RBF are examples of common kernel functions (Van Gestel et al., 2004).

The trade-off between maximising the margin and minimising the classification error on the training data is managed by the regularisation parameter (C). While a larger value of C prioritises the proper classification of training samples but may result in a

narrower margin, a smaller value of C produces a wider margin but may cause misclassification of training data. For RBF kernels in particular, the gamma parameter controls how much each training sample affects the decision border (Van Gestel et al., 2004). A decision boundary with a smaller gamma value is smoother, but one with a larger gamma value is more flexible and complex and may be more prone to overfitting (Van Gestel et al., 2004).

There are various advantages to using SVM for multi-class image classification. First of all, SVM works well with image data that has a lot of features since it performs well in high-dimensional feature spaces. Second, by utilising kernel functions, SVM can manage non-linear decision boundaries, enabling flexible modelling of complex connections in image data. Thirdly, compared to certain other machine learning methods, SVM is less prone to overfitting and has a strong theoretical base. Lastly, it has been demonstrated that SVM performs well even with relatively small training datasets, making it useful in situations with a shortage of labelled data. In general, SVM is a reliable and adaptable option with good performance and generalisation abilities for multi-class image classification applications (Foody et al., 2004).

3.3.6.2 NB

The Naive Bayes classifier is a commonly employed machine learning algorithm for classification applications, such as multi-class image classification. Its foundation is the Bayes theorem, which states that the probability of a hypothesis is determined by its observed evidence. When it comes to classification, the NB classifier determines which class has the highest probability of being the predicted class label by calculating the probability of each class given the input data (Lutfi et al., 2022).

The primary hypothesis of NB classifier is that, given the class label, the features are conditionally independent (Shi et al., 2003). This assumption simplifies the calculation of class probabilities by breaking down the combined probability distribution of the features into a product of individual conditional probabilities. This assumption frequently holds up well in practice, making the NB classifier extensible to high-dimensional feature fields and computationally practical.

Parameters of NB classifier include the prior probabilities of each class as well as the conditional probability distributions of features within each class. The prior

probabilities, which can be calculated using the training data, show the chance of each class happening in the absence of any observable features. The likelihood of detecting each feature value given the class label is described by the conditional probability distributions, which are usually modelled using probability density functions like Gaussian distributions (Lutfi et al., 2022).

Parameters of NB classifiers are commonly determined from the training data through the application of maximum likelihood estimation or other probabilistic techniques. If one were to fit parametric models to the observed feature values within each class, one could estimate the conditional probability distributions, whilst the prior probabilities might be estimated by counting the relative frequencies of each class in the training data (Lowd and Domingos, 2005).

The NB classifier has several advantages when used for multi-class image classification. First of all, NB is a good choice for image classification problems involving a large number of features since it is computationally efficient and can handle huge datasets with high-dimensional feature spaces. Second, NB minimises the chance of overfitting, particularly in situations with an absence of training data, by requiring the estimation of comparatively few parameters from the training set. Thirdly, because NB relies on the conditional independence assumption, it is resistant to duplicated or irrelevant features. This is because it basically ignores feature correlations. In multi-class image classification applications, NB offers probabilistic predictions that facilitate uncertainty estimation and model interpretability.

3.3.6.3 DT

For classification problems, such as multi-class image classification, the Decision Tree classifier is a popular machine learning method. Recursively dividing the feature space into regions, each linked to a certain class. With the help of a set of decision rules based on the values of the input features, this partitioning method enables the creation of a structure resembling a tree, with each internal node representing a decision based on a feature and each leaf node representing a class label (Van den and van Wijk, 2011).

The Decision Tree algorithm divides the training data into distinct groups by choosing the feature that best fits each phase of the partitioning process. Information gain, Gini impurity is a criterion that is commonly used to pick features. This criterion measures the degree of homogeneity or purity of the final partitions. Greater information gain indicates stronger discriminating power. Information gain is a measure of the reduction in impurity obtained by separating the data based on a specific attribute. The concept of Gini impurity evaluates the degree of uncertainty or disorder in the class distribution and seeks to reduce it by feature-based partitioning (Van den and van Wijk, 2011).

Parameters of DT classifiers consist of the maximum depth of the tree, the minimum quantity of samples needed to split a node, and additional hyperparameters that regulate the growth and pruning of the tree. The maximum depth sets a limit on the depth of the tree and helps avoid overfitting. The threshold for node splitting is based on the minimal number of samples per node; if the number of samples is less than this threshold, no more partitioning can occur. The splitting criterion and the maximum number of leaf nodes are two additional hyperparameters that affect the structure and behaviour of DT classifiers (Leiva et al., 2019).

Decision Trees are capable of learning decision boundaries and accommodating interactions between many features, in contrast to linear models, which assume a linear relationship between input features and the target class. As a result of their adaptability, Decision Trees may describe complex data patterns and relationships, making them ideal for applications involving non-linear or heterogeneous data distributions.

3.3.6.4 KNN

The K-Nearest Neighbours (KNN) performs according to the similarity principle, which states that class label of a data point is decided by the majority class of its closest neighbours in the feature space. The KNN technique maintains all training samples in memory and generates predictions based on how similar the query instance and the training examples are, rather than explicitly learning a model from the training data (Rashidi et al., 2023).

The primary method of the KNN algorithm is to calculate the distances in the feature space between each training instance and the query instance. The Euclidean distance is the most widely used distance metric (Wang, 2019). After calculating the distances, the KNN algorithm uses these values to determine which K query instance neighbours are the closest, with K being a user-defined hyperparameter (Rashidi et al., 2023).

The class label of the query instance is then chosen by a majority vote among its KNN, with the vote weighted of each neighbour by its closeness to the query instance. When it comes to multi-class classification, there are a number of ways that may be employed to separate ties. For example, one strategy is to use a distance-weighted voting scheme or to assign the class label with the highest overall frequency among the KNN (Rashidi et al., 2023).

The number of neighbours taken into account while generating predictions is controlled by the value of K, which is the primary parameter of the KNN classifier. A more specific decision boundary is produced by a smaller value of K, which may increase variance and vulnerability to data noise (Radwan Qasrawi et al., 2023). On the other hand, a higher value of K results in a smoother decision boundary and might increase adaptability against outliers, but it also increases the risk of bias introduction due to over-smoothing (Leung, 2007).

The KNN classifier has many benefits when used for multi-class image classification tasks. First of all, KNN is an easy-to-understand algorithm that works well for a variety of classification problems since it makes few assumptions about the distribution of the underlying data. Second, because KNN is memory-based and non-parametric, it can accommodate any data distributions and model complex decision boundaries in a flexible manner (Rashidi et al., 2023). Thirdly, KNN can efficiently handle high-dimensional feature spaces that are frequently encountered in image classification problems and is resilient to noisy or irrelevant features (Nimish Ukey et al., 2023).

3.3.6.5 FDA

The Fisher Discriminant Analysis (FDA) classifier is a well-known technique for supervised dimensionality reduction and classification applications, such as multi-

class image classification. It works by identifying the linear feature combination that minimises within-class variance and maximises between-class distribution. Basically, FDA aims to locate the projection of the data onto a lower-dimensional subspace with well-separated classes, allowing for effective classification (Lu et al., 2010).

FDA principally computes the within-class scatter matrix (Sw) and the between-class scatter matrix (Sb), which are the scatter matrices of the data. Whereas the between-class scatter matrix quantifies the distance between class centroids, the within-class scatter matrix gauges the distribution of data points within each class. Finding the linear transformation, sometimes referred to as the discriminant, that maximises Fisher's criterion—the ratio of between-class scatter to within-class scatter—is the main objective of the FDA (Lu et al., 2010).

Fisher's criterion mathematically aims to maximise the Fisher discriminant function, which is denoted by (Guo et al., 2022):

$$J(w) = \frac{w^T * S_b * w}{w^T * S_w * w} \quad (3.3)$$

Where:

- w is the discriminant vector (or weight vector) representing the linear combination of features.
- S_b is the between-class scatter matrix.
- S_w is the within-class scatter matrix.

The discriminant vector w is calculated by solving the generalised eigenvalue problem:

$$S_b w = \lambda * S_w * w \quad (3.4)$$

Where λ denotes the eigenvalue associated with the discriminant vector.

The FDA classifier has significant advantages in multi-class image classification tasks. Firstly, by determining the most discriminative features that divide various classes, FDA offers a useful method of dimensionality reduction. In addition to

improving computing performance, this dimensionality reduction helps lessen the negative effects of overfitting and the dimensionality curse. Second, FDA automatically takes class separability into account when extracting features, producing a feature space that best discriminates between various classes. This trait makes FDA ideal for jobs involving complex or non-linear class boundaries. FDA can also efficiently handle high-dimensional feature spaces, which are frequently encountered in image classification problems, and is resilient to noisy or irrelevant features.

3.3.7 Ensemble Models

A set of machine learning methods known as ensemble classifiers combines the predictions of several basic classifiers to enhance overall performance, including robustness, accuracy, and generalisation. Based on the principles of diversity and aggregation, these methods make use of the capabilities of individual classifiers to generate predictions that are more dependable and accurate. Several ensemble methods are frequently applied in multi-class image classification applications, including Boosting, Stacking, and Bagging (Ganaie et al., 2022). Figure 3.11 presents the different available ensemble models.

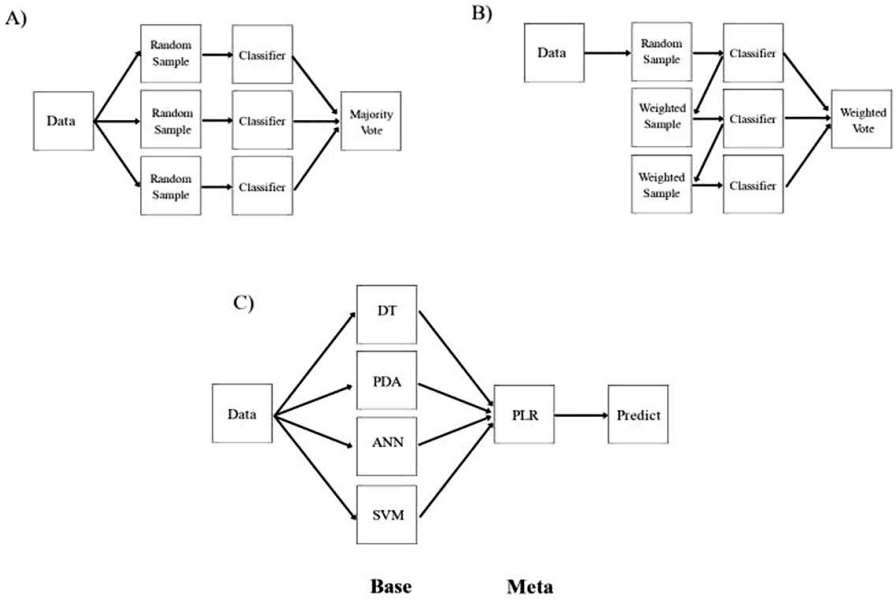

Figure 3.11 Ensemble methods. A) Bagging. B) Boosting. C) Stacking (Peterson, 2018)

Bagging, which is another name for bootstrap aggregating, is a common method used to create ensemble-based algorithms. Bagging is used to improve the performance of an ensemble classifier. Creating a series of independent observations with the same size and distribution as the original data is the fundamental concept behind bagging. Create an ensemble predictor based on the observations that outperforms the single predictor created from the original data. In the original models, bagging involves two steps: first, creating bagging samples and feeding each bag of samples to the basic models; second, developing a plan for fusing the predictions from several predictors. The combined output of the base predictors may differ because regression problems employ the averaging approach to get the ensemble result, whereas majority voting is typically used for classification issues. By training each base classifier on a different random subset of the training data, bagging creates several base classifiers. The ultimate prediction is obtained by adding together the individual forecasts, frequently by means of majority vote. Each base classifier is trained to predict the class label on its own. By averaging out the noise and errors in individual predictions, bagging contributes to a more robust and stable classifier by reducing variance and overfitting. (Ganaie et al., 2022).

Sophisticated ensemble learning methods like stacking offer a strong way to improve the robustness and accuracy of multi-class image classification systems. A meta-classifier is used in stacking to combine the predicted outputs of various base classifiers and combine these varied predictions into a final judgement. By utilising the combined intelligence of several classifiers, this method provides a comprehensive way to address the inherent difficulties of image data and enhance classification performance.

The stacking ensemble model typically consists of two main stages: the base stage and the meta-stage. Several base classifiers are trained separately on the image data in the base stage, and they all produce predictions for the target class labels. For further analysis, these predictions function as new features, or meta-features. To determine the final classification in the meta-stage, a meta-classifier is trained using both the original image features and the meta-features. By utilising the different viewpoints of the underlying classifiers, this meta-classifier gains the ability to efficiently aggregate their predictions and improve classification accuracy. A key

component of the effectiveness of the stacking ensemble model is the careful structure and selection of the basic classifiers. To ensure variation in their predictions, these base classifiers should have strengths and weaknesses that complement one another. In multi-class image classification, DT, SVM, neural networks, and CNNs are often used as base classifiers. By providing distinct skills to capture various facets of the complex image data, each base classifier allows the stacking ensemble model to take advantage of a wide range of information for classification. The stacking ensemble model has numerous benefits when it comes to multi-class image classification. First off, stacking improves the robustness of the classification model and capacity for generalisation by combining the predictions of several base classifiers. This is especially helpful in situations where noisy or unclear image data could make it difficult for individual classifiers to perform well. Second, stacking improves the ability of the model to classify images across a variety of classes by allowing it to manage complicated decision boundaries and non-linear relationships in the image data. Lastly, stacking provides scalability and flexibility, enabling the ensemble model to be adjusted to different image datasets and classification tasks, as well as the integration of several base classifiers (Mohammed and Kora, 2023).

Boosting is a well-known ensemble learning method that is essential for improving the robustness and accuracy of multi-class image classification systems. Based on the sequential training concept, boosting builds an ensemble of weak learners iteratively, with each member concentrating on the incorrectly classified samples from the previous ones. Through the iterative process of highlighting difficult cases and fine-tuning the ensemble model, the overall classification performance is gradually improved. Typically, the boosting ensemble model is made up of a series of basic classifiers, also known as weak learners or base models. These foundation classifiers are trained in a sequential fashion, with each new classifier concentrating on the cases that the previous ones misclassified. Boosting can efficiently manage complex data distributions and enhance classification performance with this adaptive training technique, especially in situations where individual classifiers may struggle with noisy or imbalanced data. Weighted training examples, in which incorrectly categorised instances are given higher weights to prioritise their correct classification in later iterations, are the fundamental building block of boosting. By focusing on

difficult cases, boosting helps to progressively lower classification errors and enhance overall performance of the ensemble model. Furthermore, the weighted predictions of all base classifiers are usually combined to produce the final prediction; the weights are established by analysing performance during training of each classifier. When it comes to multi-class image classification, boosting has many benefits. First off, boosting improves the ability of the ensemble model to reliably categorise images over a range of classes by iteratively fine-tuning it to accommodate complicated data distributions and non-linear relationships within the image data. Secondly, because boosting concentrates on the hard-to-classify cases, it is resistant to overfitting and can manage noisy or unbalanced data with ease. Furthermore, boosting is adaptable and compatible with a range of base classifiers and loss functions (Ferreira et al., 2012).

While the previous sections have explored each deep learning and classification method in detail, Table 3.6 below provides a clear and concise summary of their key advantages, limitations, and the existing gaps in knowledge. This comparative view not only supports the rationale behind selecting and combining these techniques but also underscores why further development was necessary, particularly in addressing the challenges of medical image classification in breast tomosynthesis.

Table 3.6 Comparative Analysis of Deep Learning and Classification Methods

Model/Method	Advantages	Limitations
AlexNet	• Pioneered deep CNNs in image classification • Simple and efficient architecture • ReLU accelerates convergence	• Shallow compared to newer architectures
ResNet-18 / 50	• Residual blocks prevent vanishing gradient • Enables deeper networks • Strong performance on image classification	• ResNet-18 may be too shallow • ResNet-50 computationally expensive
GoogleNet	• Efficient multi-scale feature extraction • Inception modules balance accuracy & speed • Fewer parameters than VGG/ResNet	• Complex architecture • Requires manual tuning of module parameters
DenseNet-201	• Dense connectivity promotes feature reuse • Reduces vanishing gradient • Strong hierarchical learning	• High computational cost • Requires significant memory
MobileNetV2	• Lightweight and efficient • Inverted residuals improve performance with low latency	• Lower capacity for complex feature learning • Sensitive to hyperparameter settings
VGG-16	• Simple and uniform architecture • Effective hierarchical representation	• Very high number of parameters • High memory and computational cost
SqueezeNet	• Extremely compact • Maintains reasonable accuracy	• Lower performance compared to deeper networks • Reduced expressive power
HOG Descriptor	• Captures edge, texture, and shape • Robust to lighting variations • Lightweight and interpretable	• Limited to local features • Lacks high-level semantic understanding
SVM	• Strong for high-dimensional data • Good generalization • Works well with small datasets	• Sensitive to parameter tuning • Less effective on imbalanced dataset

Naive Bayes (NB)	• Fast and simple • Good for high-dimensional data • Requires small training sets	• Strong independence assumption • Poor with correlated features
Decision Tree (DT)	• Interpretable and fast • Handles non-linear patterns • Works well with categorical data	• Prone to overfitting • Unstable with small changes in data
KNN	• Non-parametric and simple • Effective with small-scale datasets	• Computationally expensive at prediction time • Sensitive to K and feature scaling
FDA	• Maximizes class separation • Suitable for multi-class problems	• Assumes linear boundaries • Sensitive to noise and class imbalance
Ensemble Models (Bagging, Boosting, Stacking)	• Improves accuracy and robustness • Reduces overfitting and variance • Combines strengths of individual models	• Complexity in interpretation and implementation • Higher computational cost

3.4 Summary

This chapter explores the innovative effects of CAD Systems for breast cancer. CAD Systems use sophisticated learning algorithms and AI to analyse mammograms closely, helping medical professionals identify subtle early-stage abnormalities. These systems are an important additional evaluation tool that improves diagnosis accuracy and consequently promotes the well-being of patients. Furthermore, the benefits of CAD go beyond simple identification. It gives radiologists evidence-based counsel by giving statistical evaluations of lesions; at the same time, it can reduce FPs and FNs, hence reducing the need for needless medical procedures. Moreover, CAD makes it easier to track lesion changes over an extended period, allowing for more personalised treatments and better therapeutic evaluation. The adaptability of CAD systems to several aspects of breast health is demonstrated by its applications, which go beyond core tumour diagnosis and include microcalcification detection, and breast density categorisation.

The techniques employed in the integrated models in CAD systems were then discussed in the part that followed. This section provides a comprehensive overview and detailed explanation of the various approaches used in each of the five developed systems which this thesis examines. Every approach was evaluated to

have a thorough understanding of how it was implemented and how it contributed to the overall efficiency of the systems. This in-depth examination was done with the intention of giving a comprehensive understanding of the details of the methodologies used, emphasising their significance and goals in relation to the thesis.

Chapter 4 Review of Digital Breast Tomosynthesis Research

4.1 Introduction

This chapter explores the DBT research that has been done. A comprehensive review of the literature in the field of DBT reveals a heterogeneous research environment aimed at improving breast imaging efficacy. A primary research focus is on continuously enhancing image quality using cutting-edge methods that address artefact reduction, contrast-to-noise ratio, and spatial resolution in DBT scans. Researchers like Gao et al. (2020), Gao, Fessler and Chan (2021), Su et al. (2021), Siti Noraini Sulaiman et al. (2022), Syafiqah Aqilah Saifudin et al. (2022), Mota, Mendes, and Matela (2023), and other academics have made some contributions in this area of study. Reducing patient exposure while maintaining diagnostic precision is paramount, as demonstrated by the work of eminent experts studying radiation dose optimisation, including Tsutomu Gomi et al. (2022) and Ajay Kumar Visvkarma et al. (2022). These investigations also thoroughly study the impact of different acquisition conditions on the quality of the images. The comparative evaluations of the literature assess clinical performance of the DBT and diagnostic accuracy, especially when compared to standard mammography. This helps to clarify the efficacy of the technology for a range of breast densities and lesion types.

Integration with other imaging modalities, such as mammography, appears to be a focus of exploration and development by certain researchers, such as Wang et al. (2021) and (R V, R and A P, 2021). The goal is to uncover potential benefits that could boost overall breast cancer detection rates. The field of 3D image analysis and computer-aided detection is also well covered in the literature, which presents the creation and assessment of sophisticated algorithms intended for automated lesion identification and characterisation in DBT scans. A recurrent feature in the research is the integration of AI and machine learning methods with CAD software, highlighting a dedication to utilising state-of-the-art technologies to enhance diagnostic results. In addition, the literature carefully studies how DBT affects radiologist workflow and interpretation time, to find optimisation tactics that would improve therapeutic effectiveness. Additionally, research on the cost-effectiveness of the DBT, accessibility tactics, and long-term effects on patient quality of life, survival

rates, and detection rates of breast cancer are all included in the literature. The summary of current research essentially highlights a shared commitment to improving DBT, to improve patient care and breast cancer detection.

Over the course of this doctoral research, a thorough analysis was conducted that covered several aspects, such as the previously mentioned improvements in image quality, detection methods, and radiation dosage reduction. Notably, the current thesis does not provide results or information regarding these domains. Specific investigations were conducted to examine various aspects of the research on DBT scans. This study primarily focuses on the use of CAD Systems and deep learning approaches for the classification of DBT images.

This segment comprises two distinct subsections. A thorough literature assessment of computer-aided detection systems used in DBT scans is presented in the first subsection, along with an analysis of deep learning applications designed for classification of DBT scans. The subsection that follows offers a thorough analysis of the systems that were utilised when combined with the BCS-DBT dataset (Buda et al., 2020). Notably, this dataset functioned as a foundational dataset for the research carried out within the scope of this PhD study.

4.2 CAD Systems and deep Learning applications in Digital Breast Tomosynthesis

A hierarchical model of latent bilateral feature representation was presented by Kim et al. (2016) as an approach of classifying masses according to the asymmetry of the left and right breasts. The Samsung Medical Centre provided the researchers with a dataset that comprised 160 reconstructed volumes from 40 different individuals. Of these volumes, 86 had at least one biopsy-proven malignant tumour, whereas the remaining 74 were found to be normal. Volume registration of the DBT main and lateral images was the first step in the process. The bilateral feature representation was then extracted using a 3D-CNN after the Volume of Interest (VOI) transform. For mass classification, the characteristics from the lateral view and main view VOI were combined and fed into a fully linked layer. Based on the Area Under the Curve (AUC) values of 0.847 and 0.826, respectively, the suggested model outperformed the hand-crafted feature classifier. However, the small dataset limits the generalizability

of the findings, and the approach focuses solely on bilateral asymmetry without accounting for broader lesion variability. Additionally, the model lacks external validation, which further restricts its applicability in diverse clinical settings.

Fotin et al. (2016) compared the identification and classification of two types of breast cancer abnormalities from 3D (DBT) images using a DCNN with a conventional method. To assess the performance of detection, a unique test set including 344 reconstructions from DBT was employed. 328 suspicious and 115 malignant soft tissue densities, including masses and architectural deformities, were included in these reconstructions. They were produced using the GE SenoClaire 3D and an iterative reconstruction technique. The traditional method was to manually extract features from the Region Of Interest (ROIs) and then apply them to an ensemble of boosted DT. The DCNN approach, on the other hand, resizes the ROIs to 256×256 and submits them to a DCNN that is nearly identical to that of AlexNet for the purpose of abnormality detection and classification. When switching from the conventional to the deep learning approach, researchers observed that the sensitivity of the ROIs increased from 83.2% to 89.3% for suspicious ROIs and from 85.2% to 93% for malignant ROIs. This implies that deep feature learning has significant potential for a range of medical image analysis tasks and is very helpful for the interpretation of DBT data. However, the ROI resizing process may lead to a loss of spatial context, and the study remains limited to the detection of soft tissue densities without addressing other abnormality types. Furthermore, the approach was not evaluated on diverse datasets, raising concerns about its generalizability.

Rodriguez-Ruiz et al. (2017) utilised a CNN model to classify calcification by utilising reconstructed images with different DBT reconstruction algorithms, filtered back projection (FBP), and FBP with iterative optimisations (EMPIRE). The model used a dataset that includes 2,071 patient studies of DBT performed at the institution of the author between December 2014 and December 2015 during standard clinical examinations in accordance with established protocols. This collection includes 30 instances that were biopsy-confirmed as benign, 40 cases that were confirmed as malignant, and 30 more cases that were identified as normal. The model was inspired by OxfordNet and included one batch normalisation layer, four convolutional layers, one maxpooling layer, and three FC layers, with input dimensions of 29 ×29

×3 . The AUC of the model trained with images reconstructed using the FBP algorithm was 0.857, and the AUC of the model trained with images reconstructed using the EMPIRE algorithm was 0.880. To sum up, the EMPIRE reconstruction algorithm has been shown to have superior contrast and image quality, fewer artifacts, and improved visibility of calcifications as judged by human observers, as well as improved detection capability in deep-learning systems when compared to FBP. This could lead to improved clinical performance of radiologists and more accurate computer detection systems using deep learning. However, the comparative analysis was limited to only FBP and EMPIRE algorithms, with a relatively small number of malignant cases in the dataset. Moreover, the lack of validation set limits the generalizability of the findings.

In Zhang et al. (2018), the authors created a range of CNN models, including AlexNet and ResNet50, to classify 2D mammograms and 3D tomosynthesis scans. The dataset they used was composed of 3018 negative and 272 positive exams. They implemented data augmentation and transfer learning techniques in their research. Out of the proposed models for 2D mammograms, 2D-T2-Alex achieved the highest auROC of 0.73. This model extracted features from a pre-trained AlexNet and then employed a shallow CNN to classify the exams as either positive or negative. The shallow CNN had one convolutional layer with a kernel size of 1×1 and 256 depth of the filter, followed by two 1024 FC layers. For 3D tomosynthesis classification, 3D-T2-Alex had the best performance, with an auROC of 0.6632; similarly, to the 2D mammogram images, transfer learning using AlexNet was able to improve the performance of the 3D tomosynthesis classification models. However, the 3D DBT models demonstrated relatively low AUC values, indicating weak performance in generalizing from 2D to 3D. Additionally, the transfer learning approach was not fully optimized for the unique characteristics of DBT data.

The suggested model for identifying DBT images as abnormal or normal by Samala et al. (2018) employed two steps of transfer learning. The pre-trained AlexNet model was initially adjusted with FFDM images after being trained with ImageNet. Afterwards, the model was trained using DBT z-Stack images using the previously trained weights from FFDM images, and the CNN was used to extract features. A feature selection method was then utilised, followed by a random forest classifier.

AlexNet was originally fine-tuned using 19632 augmented ROI patches from 2454 mass lesions. 9120 ROI patches from 228 mass lesions from DBT z-Stack images were employed in the second round of training the model. To assess the performance of the proposed model, they changed the parameter sizes of the model and calculated the AUC. They employed network pruning, which resulted in a reduction of 87.2% in trainable neurons, 34.4% in trainable parameters, and 95.5% in multiplications and additions, in order to further simplify the model. Findings from the research indicate that the pruned CNN can attain an AUC of 0.90, similar or surpassing the AUC of 0.88 of the non-pruned CNN. The outcomes demonstrated that, although having fewer parameters, the pruned DCNN outperformed the original DCNN in terms of accuracy, indicating that evolutionary pruning may be a helpful technique for enhancing the performance of pre-trained DCNNs for medical image analysis. However, the patch-based approach may lose global lesion context, the diversity of data augmentation was limited, and the model remained dependent on manually annotated ROI patches.

A detailed comparison of two different frameworks that both focused on supervised classifier architectures was carried out by Bevilacqua et al. (2019). Using hand-crafted morphological and textural features extracted from each ROI, the original architecture used a feature-based technique to feed optimised Artificial Neural Network (ANN) classifiers. The second framework, on the other hand, extracted different feature sets using different CNN models to evaluate classification performance using non-neural classifiers that used automatically generated features. The research employed a private dataset that included 39 DBT exams. Morphological and textural features were extracted after input image processing and segmentation in the ANNs framework. The activation of the final layer of each CNN model was utilised as an input for a different learner when it came to feature extraction in the second framework. A variety of CNN models that had been trained beforehand as well as non-neural classifiers (VGGG-F, VGG-M, and VGGS) were evaluated. These models performed better when combined with the KNN classifier. The results showed that the second framework, which employed features automatically extracted by CNN architectures, performed better in the final evaluations than first framework in terms of accuracy, specificity, and sensitivity. With VGG-S and KNN, the second framework achieved an accuracy of 93.26%.

However, the study was based on a very small dataset, used outdated hand-crafted feature comparisons, and lacked testing on any external benchmark datasets.

By investigating techniques to integrate the 2D slice images in the reconstructions, Zhang et al. (2019) expanded on their previous work and investigated different methods to incorporate complete DBT volumes into pre-trained models created for 2D data. They specifically suggested a model that had two stages of fusion, early and late. Two classification strategies have been investigated in the early stages of fusion. The feature extractor creates a single pseudo-2D mammography by averaging all the z-slices. The classifier then uses the feature map as its only input and outputs the expected label. The second method involves fusing the series of images using dynamic image networks, which input z-slices and produce a dynamic image. Features are initially retrieved for late fusion, and then a single feature map is created by pooling those features. On DBT, three different pooling types—minimum, average, and maximum—were investigated. The authors of this study compared the early fusion stage and late fusion stage performances to their Three-Dimensional AlexNet (3D-AlexNet) model that they had previously developed in a prior study. Additionally, they examined the outcomes of the late fusion stage while employing various feature pooling techniques. Using max pooling and AlexNet transfer learning, the late fusion stage achieved the best performance. Their proposed approach generates an auROC of 0.854, a 28.80% improvement of auROC of 0.663 over their previous model. However, the study reported that space-to-channel encoding using adjacent z-slices reduced performance, and early fusion methods underperformed overall. Additionally, the pre-trained models used may not effectively capture DBT-specific spatial patterns, and end-to-end retraining to improve this was computationally infeasible.

Mendel et al. (2019) investigated how various breast cancer screening techniques affected performance when deep learning methodology with transfer learning strategy was used. Using pre-trained DCNN VGG-19 and SVM, they separately retrieved and categorised the features of 78 mammography lesions from a total of 76 patients. Biopsy results for the 78 lesions revealed 48 to be either high risk or benign, and 30 to be malignant. They investigated the feasibility of employing feature extraction to train pre-trained CNNs for the classification of malignant from benign

breast lesions on FFDM, synthetic 2D images, and DBT key slice images. For each of these modalities, features were taken out of the max pooling layer of the VGG19 convolutional network and passed via an average pool layer to minimise the number of features. Initial feature extraction was followed by feature dimension reduction. After feature extraction and reduction, a non-redundant collection of useful features was found using a leave-one-out stepwise feature selection process. One training example was excluded from each cycle of stepwise feature selection to identify such a feature set over the training data. For each breast imaging modality (FFDM, Synthesised 2D, and DBT) and each view (CC, MLO and merged CC and MLO), the AUC was calculated for the categorisation of malignant and benign lesions. In the task of lesion characterisation, the synthesised 2D image performed best in both the CC view and the MLO view (AUC=0.81, SE=0.05, and AUC=0.88, SE=0.04, respectively). Soft voting was used to combine the CC and MLO data for each lesion, and the DBT key slice image performed the best (AUC=0.89, SE=0.04). Based on their research, it is stated that, as compared to conventional Full-Field Digital Mammography (FFDM), DBT excelled at enabling pre-trained CNNs to maximise their strength as feature extractors. However, their work is limited by a small dataset of 78 lesions and lacks end-to-end model training, relying instead on static feature extraction from pre-trained VGG-19. Additionally, the study does not include multi-site validation or account for clinical variability, which may impact the generalizability of its findings.

Using radiomics for a thorough examination of radiological images, (Sakai et al., 2019) developed an automated classification system for diagnosing breast lesions in DBT images. The University Pierre and Marie CURIE (UPMC) Breast Tomography and Aoyama Hospital in Japan provided the tomosynthesis dataset used in this study, which included 20 cases of benign lesions and 31 cases of malignant lesions. The authors computed 70 radiomic characteristics, including lesion morphology, spicula presence, and textural information, and defined an analysis area centred on the lesion. By feeding the generated radiomic characteristics into four classifiers—SVM, RF, NB, and MLP—accuracy was evaluated. SVM obtained an accuracy of 55% in detecting benign tumours and 84% in detecting malignant tumours.

To classify benign and malignant mammograms, Liang et al. (2020) used a hybrid CNN that integrates both 2D and 3D CNNs and depends on both FFDM and 3D DBT. By combining pretrained deep learning models as the basis for feature extraction before feeding the results into their 2D and 3D convolutional layers, the authors employ this 2D-3D ensemble technique. First, they utilised data pre-processing, which involves reconstructing 2D dynamic images from DBT slices. Next, for feature extraction, the recovered 2D dynamic images and FFDM images are fed into four deep learning models (AlexNet, ResNet, DenseNet, and SqueezeNet). Afterward, three classifiers retrieve the extracted features (DM Classifier, DBT Classifier, and DM-DBT Classifier, which uses DM only, DBT only, and DM and DBT data, respectively). The new multimodal model, fuses extracted features by weight-sharing and obtains an AUC of 0.97 when both mammographic images are ensembled. When trained on independent modalities, the model achieves 0.87 and 0.72 AUC on DM and DBT, respectively. However, their system relied on a limited dataset of just 51 cases, which restricts the robustness and generalizability of the results. Additionally, the study used traditional radiomic features without incorporating deep learning approaches or validating across diverse external datasets.

Zheng and Mo (2020) developed an End-to-End multi-scale multi-level features fusion Network (EMMFFN) model for DBT mass classification. The private dataset of the study, which included 471 masses from 441 patients and 927 views, was collected in Marlborough, Massachusetts, in the United States. Three distinct images of the breast mass were taken from ROIs: the gross mass, overview, and mass background. Subsequently, these representations were concurrently input into the EMMFFN model, producing three sets of feature maps. The EMMFFN model, which is composed of up of three sub-models that have been improved by DensNet-121, combines these three feature maps at the feature level to provide the final prediction. The results showed that in terms of breast mass classification, the EMMFFN model achieved an AUC of 85.09%. However, the performance of their system was only validated on a private dataset, limiting generalizability. Moreover, the study lacked comparison with other state-of-the-art fusion architectures and did not explore clinical interpretability or external validation.

The authors of Singh et al. (2020) suggested a technique that focuses on how to incorporate DL models created for FFDM examinations to DBT exams. Initially, during training, images were augmented with random horizontal and vertical flipping. After using FFDM images to train a 29-layer ResNet-based model, the model was subsequently trained with 2D maximum intensity projections (MIP) of DBT images. To better match the FFDM images that were initially used to train the algorithm, MIP images of DBT are transformed using Histogram Matching (HM). To fine-tune the base model trained on FFDM images for usage with the original or histogram-matched MIP images, two techniques were applied. Only the last completely linked layer was retrained in the first strategy. This is known as the traditional fine-tuning approach. A variant of the SpotTune method, which determines the best layer to fine-tune for each occurrence of target data, was used for the second approach. When evaluated on image patches derived around reported observations, their method achieved AUCs of 0.9 for FFDM and 0.85 for MIP images, compared to 0.75 when tested directly on MIP images. Their system primarily relied on 2D Maximum Intensity Projections (MIP) from DBT, which may not fully preserve the 3D contextual information necessary for accurate diagnosis.

Li et al. (2020) investigated the efficacy of DCNNs embedded with different transfer learning algorithms for mass categorisation utilising DBT and FFDM in a study. They investigated a total of 1854 2D/3D ROIs (FFDM or DBT) in the dataset, including 927 mass ROIs (665 benign and 262 malignant) and 927 normal ROIs. To evaluate the capabilities of DBT and FFDM in mass classification inside the DCNN framework equipped with transfer learning and to investigate a viable combination approach of DBT and FFDM in improving classification performance, three investigations were undertaken. The purpose of this mass classification model was to classify images as cancerous, benign, or normal. The first study assessed the classification performance of VGG-16-based Two-Dimensional Deep Convolutional Neural Networks (2D-DCNNs) models trained by DBT and FFDM with or without transfer learning using the natural image database ImageNet. In the second study, researchers investigated several ways for merging DBT and FFDM (where both modalities are provided for each patient) in training a VGG-16-based 2D-DCNN. The 2D-DCNN can be integrated with either a Single Transfer Learning (STL) or a double transfer learning (DTL), or a mix transfer learning (MIX) of the DBT&FFDM. The

classification performance of an 11-layer 2D-DCNN and an 11-layer Three-Dimensional Deep Convolutional Neural Networks (3D-DCNN) trained from scratch using DBT was also assessed in the third investigation. The most successful model, DTL DBT, achieved mean AUC (Standard Deviation) values of 0.910 (0.012) for malignant tumours, 0.954 (0.003) for benign tumours, and 0.984 (0.004) for normal tissues. According to the experimental findings, transfer learning improved model performance for both DBT and FFDM, and when using transfer learning, the DBT-based model generally surpassed the FFDM-based model. Additionally, classification accuracy can be effectively increased by incorporating the two picture modalities DBT and FFDM during model training. Their system faced limitations due to the use of relatively small and imbalanced datasets, which can affect the model's generalisability and robustness. Moreover, while multiple transfer learning strategies were explored, the study lacked external validation on multi-institutional datasets and did not assess real-world clinical applicability.

Fan et al. (2020) introduced a Three-Dimensional Mask (3D-Mask) RCNN-based breast mass segmentation model that utilised a ResNet-Feature Pyramid Network (ResNet-FPN) for extracting multiple feature pyramid scales. Top-down and bottom-up features were infused in different scales by FPN. To create candidate bounding boxes from the input image, a region proposal network (RPN) was then employed. With the use of a classifier network and bounding-box regression, the detection branch carried out mass detection for each proposed ROI to determine the location for the boxes. To predict a segmentation mask from each ROI using a Fully Connected Network (FCN), the mask branch used location of the data from the feature maps. Comparisons between the proposed 3D-Mask RCNN, 2DMask RCNN, and Faster RCNN were undertaken. The 3D-Mask RCNN delivered a sensitivity of 90% for breast-based mass detection at 0.83 FPs/breast, outperforming the 2D-Mask RCNN and Faster RCNN, which delivered a sensitivity of 90% with 1.24 and 2.38 FPs/breast, respectively. The findings imply that both the entire dataset and subsets with various attributes benefit from the 3D-Mask RCNN CAD framework over 2D-based mass detection models. Their system presented a 3D-Mask R-CNN framework for DBT mass detection, but the study's main limitation lies in its reliance on a relatively limited dataset, which may affect the model's generalizability.

Additionally, the comparison with 2D-based methods lacked extensive statistical analysis and validation across diverse imaging conditions and institutions.

A two-level framework for breast cancer categorisation was implemented by Aswiga et al. (2021). To train and categorise the target DBT dataset, the first level constructed a simple Multi-Level Transfer Learning (MLTL) framework utilising the information gained from generic non-medical image datasets (ImageNet) and the mammography dataset (MIAS Dataset). The target, intermediate, and source domains contribute their MLTL framework. The intermediate domain, which consists of the mammography dataset, receives the essential common features from the source domain, which is a widely accessible non-medical images dataset. The DBT dataset is then classified once the features of the mammography dataset are transferred to the target domain. The performance of the MLTL framework is enhanced by the suggested second-level framework, Feature Extraction-based Transfer Learning (FETL), employing the following feature extraction approaches. The CNCF fusion algorithm is the first approach used to merge the high-level and low-level features of the target domain images. The second approach extracts a collection of texture features from the target domain images using Gray-Level Co-Occurrence Matrix (GLCM). The third approach employs a multi-input perceptron algorithm to extract features from both the target domain images and the patient reports. The first framework achieved an auROC of 0.81, whereas the second framework achieved auROC curves of 0.89,0.88, and 0.89 when employing CNCF fusion approach, GLCM-based feature extraction algorithm, and MIP algorithm, respectively. Their system proposed a two-level transfer learning framework for DBT classification, yet the study is limited by the small dataset size and lack of external validation, which constrains the model's generalizability. Furthermore, while the feature extraction methods improved performance, their integration increased model complexity, which may hinder real-time clinical applicability.

An automatic detection approach was developed in the Ricciardi et al. (2021) study to categorise the existence or lack of mass lesions in DBT. Three distinct DCNN architectures functioning at the image level (DBT slice) were evaluated: two advanced pre-trained DCNN architectures (AlexNet and VGG19), one customised by transfer learning, and one developed from the ground up (DBT-DCNN). Two

separate datasets from various hospital radiology departments were used to assess the performance of these DCNN-based algorithms. The DBT slice images received data augmentation and normalisation as part of the pre-processing steps. The accuracy, sensitivity, and AUC values were assessed on both datasets using receiver operating characteristic curves. Moreover, the location of the lesion in the DBT slice was displayed using the Grad-CAM technique. AUC, sensitivity, and accuracy values obtained from the analysed DCNN are in line with the most widely reported outcomes in the field. To be more precise, the DBT-DCNN network had an AUC of 0.89 ± 0.04, accuracy of 90% ± 4%, sensitivity of 96% ± 3%, and the F1-score of 0.93 ± 0.03. Based on the Grad-CAM output, it is concluded that using this method could be an efficient method for locating ROI on slides that would be used to train the DCNN and enhances the localization process of masses. Their developed DCNN architectures lacked a thorough comparison with newer state-of-the-art deep learning methods and did not explore the integration of 3D contextual information across slices.

4.3 State-of-the-Art Systems Utilising BCS-DBT for Breast Imaging

An innovative technique to combine detection possibilities from many models with the fewest possible FPs was put forth by Shoshan et al. in (2021). FPN and ResNet backbone architecture were deployed to build RetinaNet-based object detectors. For training and testing, a dataset from DBTex (DBTex Challenge) and one in-house dataset were both used. Predictions for bounding boxes in this model were subjected to numerous phases for a particular 3D mammography volume, including the collection of the top three prediction boxes from each slice across all detector models for a given volume. From there, all scores were normalised linearly according to each model. Afterwards, slice classifier weighting (CLS), which was trained without complete localisation information, adds the classifier model, and computes a score per slice. Thereafter, bounding boxes were size-filtered and boxes with a diagonal larger than 800 pixels were removed. The results of several detector model predictions were then represented by a heat map. The top 6 bonding boxes were then picked for evaluation using Non-Maximal Suppression and Rank Based Score Modification (RBSM). Finally, 3D prediction boxes are generated. Two performance metrics were calculated in their results. The primary metric, which only includes views with a biopsied discovery, is the average sensitivity for 1, 2, 3, and 4 FPs per

DBT view. The average sensitivity for two FPs across all views becomes the secondary metric. In the DBTex challenge, they came in second place with a model that recorded a primary metric of 0.910 and a secondary metric of 0.904, while the highest model recorded a primary metric of 0.912 and a secondary metric of 0.912, and the third-place model recorded a primary metric of 0.853 and a secondary metric of 0.868. Their sophisticated multi-model ensemble detection approach for DBT, but the method involved complex post-processing steps that may hinder scalability and real-time application. Furthermore, the reliance on slice-level classifiers without full 3D lesion context limits interpretability and may reduce sensitivity in challenging diagnostic cases.

To classify DBT volumes, Tardy and Mateus (2021) developed and evaluated a novel slab-based classification system. The slabbing technique is used in conjunction with a Multiple Instance Learning (MIL) classifier in this approach, which depends on local summarisation of DBT slices and only requires volume-wise labels for training. The authors showed that the approach can maintain mammography knowledge while optimising performance on DBT data through transfer learning trials. This unique attribute allows the classifier to be trained simultaneously on DBT slabs and mammograms, removing the requirement for modality-specific fine-tuning. The BCS-DBT dataset (Buda et al., 2020) and a Proprietary Multi-Vendor Mammography dataset (called PMV-MG) served in the trials for performance consistency evaluations and network pretraining. On the BCS-DBT test set, the model achieved an AUC of 0.73. To summarise, the slabbing strategy minimises the computing complexity of the classifier. However, their system introduced a slab-based classification framework using Multiple Instance Learning (MIL), which reduced computational complexity but oversimplified 3D spatial information by summarizing slices.

CNNs were trained on the BCS-DBT dataset (Buda et al., 2020) by Fogleman, Otsap, and Cho (2021), which allowed the CNN to discriminate between two classes: normal and abnormal (benign or malignant). There were two different approaches used by the researchers. Two deep learning models, VGG-16 and InceptionV3, were used in the first method, which involved transfer learning. Prior to and after the application of image augmentation techniques, the VGG-16 model was trained using

Adam optimisers and Stochastic Gradient Descent (SGD). With 8192, 2048, 512, and 128 nodes, the authors then added four dense layers. The second approach utilised the InceptionV3 model and introduced a dense layer with 128 nodes and cutting layers. The first model achieved an optimal accuracy of 64%, while the second model demonstrated the highest accuracy of 94.9% in binary classification.

Nogay, Akinci, and Yilmaz, (2021) classified DBT data using five traditional pretrained DCNN, which are ResNet-18, AlexNet, GoogleNet, and ShuffleNet. Using transfer learning techniques, several layers of pre-trained DCNN models were modified to comply with the aims of the study. This allowed the models to adjust to their new context. To enable quicker adaptation to the changing DCNN models, new weights were assigned to the newly developed layers in the five pre-trained DCNN models, while the weights of the existing layers stayed the same. The BCS-DBT dataset (Buda et al., 2020) provided the dataset used in this study, which was divided into two subsets for the classification. The first subset facilitated binary classification (Cancer+Actionable and Benign+Normal), while the second subset enabled quadruple classification (Cancer, Actionable, Benign, and Normal). Each convolution layer in every DCNN was associated with the ReLU activation function. In the test findings, accuracy rates varied between 65% and 75% for the first subset and between 66% and 86% for the second subset. According to the results, AlexNet demonstrated the maximum accuracy at 75% for quadruple classification using the second subset, while ResNet-18 earned the highest accuracy at 86% for binary classification using the first subset. Their work was limited by modest classification accuracy and reliance on shallow architectures without domain-specific optimization.

For detecting biopsy-proven breast lesions on DBT, Hossain et al. (2022) suggested a multi-depth level convolutional model utilising non-biopsied samples. They used the dataset from the DBTex challenge stated above (DBTex Challenge). First, they generated 2D slices from DBT volumes including benign lesions and breast lesions that had not yet undergone biopsy. The slices were then joined with slices that were immediately adjacent to the lesion to create 2.5-dimensional (2.5D) images of the lesion by allocating them to the RGB colour channels. To boost the amount of training samples, they augmented the 2.5D images of slices that were immediately next to the lesion centre. They employed the YOLOv5 algorithm as their basis

network for lesion detection. To find actionable FPs in non-biopsied images, they trained a baseline algorithm (medium-depth level) using biopsied samples. The baseline model was then adjusted using the augmented image set (actionable FPs added). The DBT volume was processed slice-by-slice to estimate bounding boxes in each slice, and they integrated the bounding boxes by linking them along the depth using volumetric morphological closing for lesion inferencing. They repeated the technique above to train a second model (big) with deeper-depth levels. Finally, they combined the medium and large detection models to create an ensemble method. They assessed their techniques using the free-response operating characteristic curve. They only calculated mean sensitivity per FPs per DBT volume for views that were biopsied; for all views, they calculated sensitivity at 2 FPs per Image (FPI). Their ensemble model achieved a mean sensitivity of 0.786 FPs per DBT volume (the primary metric from the study above) and a sensitivity of 0.743 at 2FPI (the secondary metric from the study above) on the DBTex independent test set. Their multi-depth convolutional model leveraging non-biopsied samples, but the reliance on false positives for model tuning could introduce bias and limit robustness. Moreover, the use of 2.5D representations may inadequately capture volumetric context, and the ensemble model was not benchmarked against other state-of-the-art 3D detectors.

A unique deep neural network capable of learning from and making predictions from high-resolution 3D medical images was proposed by Park et al. (2022) and is known as the Three-Dimensional Globally Aware Multiple Instance Classifier (3D-GMIC). By first identifying the regions of interest with a low-capacity sub-network and then selectively applying a high-capacity sub-network to the regions of interest while avoiding processing duplicate information from nearby slices, 3D-GMIC effectively concentrates its computation to the small subset of important regions. This study considered two datasets, the first of which is the internal dataset, which includes 99,862 tests from DBT, FFDM, and synthetic 2D scans. The second dataset comes from Duke University Hospital (Buda et al., 2020) and is a subset of it that was made available as the training dataset for the DBTex challenge1 (DBTex Challenge). The global module and the local module are the two subnetworks that make up their suggested design, known as 3D-GMIC. The proposed global module expands GMIC to 3D data. CNNs are employed within the low-capacity global network to handle 2D

input images. The global network is parameterised as ResNet-22, much like in the original GMIC. Each slice of a 3D image is subjected to a distinct parallel application of the global network by 3D-GMIC. The global network initially extracts the hidden representation for each slice in the input 3D image. A semantic segmentation layer is then used to convert the hidden representation into saliency maps. According to the saliency maps, they chose the most crucial areas for the 3D-GMIC local module. They specifically choose square patches that match the high saliency map values. To create the final class predictions, they average the predictions from the global and local modules of 3D-GMIC. By minimising binary cross entropy (BCE) losses for the predictions from the two stages, the loss function employed in the training of the 3D-GMIC is trained end-to-end. Their model classified DBT images with malignant findings to the external dataset from Duke University Hospital with an image-wise AUC of 0.848 (95% CI: 0.798-0.896) and classified DBT images with benign findings with an AUC of 0.741 (95% CI: 0.697-0.785). They introduced the 3D-GMIC model, which effectively localized key regions in high-resolution 3D DBT images, but the reliance of the model on patch-level inference may miss global contextual information.

Bai et al. (2022) developed a GCN model to construct a novel model for more exact identification of malignant 3D mammography images. Two datasets were utilised in this work to develop and assess the effectiveness of the suggested model. The first dataset, the BCS-DBT dataset (Buda et al., 2020), is available to the public, and the second dataset is a private dataset. To train and evaluate the suggested model, they merged both datasets and employed 402 3D mammograms (169 cancer and 233 normal) z-Stacks in total. Multi-scale Graph Convolution Network (MGCN), a suggested DBT classification GCN model, combines two techniques: spatial-based self-attention pooling Graph Convolution Network (GCN) and graph representation. The underlying features between slices and the features of the slices themselves were quickly learned using the MGCN model. During their trials, they compared the accuracy, sensitivity, precision, F1 and AUC of their model to that of baseline models such as 3D ResNet, ResNet-Vote, Two-stream, and Spatial ResNet. With scores of 0.84, 0.86, 0.84, 0.83, and 0.87, respectively, the results demonstrate that their suggested model outperforms all baseline models in terms of accuracy, precision, sensitivity, F1, and AUC. GCNs exhibit interesting results with the right architecture,

even though there are currently no standards for representing images as graphs. Their system demonstrated strong performance but relied on graph-based representations that are not yet standardised in medical imaging. Moreover, the fusion of private and public datasets without detailed harmonisation procedures could limit reproducibility and generalisation of the findings.

In collaboration with a SVM classifier for a DBT dataset, Hassan et al. (2022) suggested a breast tumour classification approach based on examining and comparing the effectiveness of diverse and the most cutting-edge deep learning classification models. They focused on the capacity to classify cancers in unseen DBT medical images using transfer learning from non-medical images by employing the fine-tuning technique to increase classification accuracy. Tumour patch cropping and data augmentation are the two processes that make up the data preparation step in this study. They utilised the DBTex challenge (DBTex Challenge), a DBT images dataset that is openly accessible. In terms of ROI selection for the tumour patch cropping stage, tumours are centre cropped and extracted from the annotated DBT images. To accommodate the input of deep learning networks during the feature extraction stage, all the identified ROIs are subsequently scaled to the same dimensions of 224x224. There have been two stages to the radiomics extraction procedure. In the first stage, a transfer learning strategy is utilised to directly extract features from the medical DBT images and train one of the classifiers in the classification heads. These CNN models were previously trained on non-medical images. The training set of 246 DBT images is used to fine-tune classification CNN models that have been pre-trained on non-medical images in the second stage. They examined the effectiveness of many of the most cutting-edge deep learning models, including AlexNet, VGG, ResNet, WideResNet, SqueezeNet, and EfficientNet, for feature extraction stages. Two classification scenarios, the first utilising end-to-end deep learning classification using a FC layer and the second using an SVM classifier, were applied to classify the generated tumour radiomics from the feature extraction step. Their tests revealed that employing end-to-end deep learning and SVM to categorise radiomics obtained from the fine-tuned AlexNet model provided the best classification accuracy of 80.43% and a 71.74%, respectively. Their system relied heavily on transfer learning from non-medical image datasets, which may not

capture domain-specific features relevant to DBT images. Additionally, the dataset used was relatively small.

A CNN of three Convolutional layers, three Pooling layers, one Flatten layer, and two Dense layers was designed by Adhikesaven et al. (2022). The first convolutional layer, called Conv2D, included 16 filters, each measuring 3 by 3. The rectified linear activation function (ReLU), a piecewise linear function, was applied by this convolutional layer, which required that the input images have size of 300 x 300 pixels. The number of filters in the CNN was systematically increased throughout layers by the researchers to improve the abilities of the network to extract information from image input as it passed through the layers. A total of 250 images were obtained by randomly selecting 125 scans that were malignant and 125 scans that were non-cancerous from the BCS-DBT dataset (Buda et al., 2020). It was decided that this sample size would be adequate for building a reliable training and testing dataset. Using a split ratio of 60/20/20, the data were divided into three subsets: training, validation, and testing. In binary classification, the model showed a remarkable 97.25% accuracy. However, their system used a small and imbalanced dataset (250 images), which limits the statistical robustness and generalisability of their findings. Furthermore, the model architecture was relatively simple and lacked comparison with more advanced deep learning approaches, leaving its competitive performance uncertain.

A model consisting of two identical parallel CNNs with shared weights, known as twin networks, and a distance learning network was presented by Bai et al. (2022). This model of shared weight twin networks, often known as the Siamese model or FFS-CNN, was created with the explicit objective of obtaining intra-image feature representation from two images that were viewed in the current and prior years. After extracting inter-image attributes from the matched images, the model uses a distance learning network to forecast how similar breast tissues will be. The authors conducted comparative analyses against several baseline models, including feature fusion models like a vanilla Siamese network, a Longitudinal LSTM model (LLSTM), and well-known deep learning models like VGG and ResNet, to evaluate the performance of their proposed model, FFS-CNN. Four datasets were utilised in the study, three of which were for training and one for testing. Included in these datasets

were the Chinese Mammography Database (CMMD) (Cui et al., 2021), the Digital Database for Screening Mammography (DDSM) (Clark et al., 2013), BCS-DBT (Buda et al., 2020), and a proprietary dataset from the Radiology Department at the University of Connecticut Health Centre (UCHC). Performance-wise, their proposed model outperformed the baseline models, attaining notable measures such as 92% accuracy, 93% sensitivity, 91% precision, 91% specificity, 92% F1 score, and 0.95 AUC. However, their reliance on matched longitudinal data limits its applicability in real-world settings where such data may be unavailable. Additionally, while performance was strong, generalisability across diverse datasets and imaging conditions was not fully evaluated.

A deep learning model specifically designed for the classification of lesions into benign or malignant categories was developed by Mendes et al. (2023). The model was built using a foundational framework based on previous research by Muduli et al. (2021), incorporating minor modifications to the FC layers architecture and regularisation technique. A total of 77 volumes from DBT—of which 38 were malignant and 39 were benign—were carefully chosen from the BCS-DBT dataset (Buda et al., 2020). Nine slices total from each volume were included, four before and four after the lesion-slice, one of which showed the lesion at its most visible stage. Three systematic applications of well-established data augmentation techniques (rotation, translation, and mirroring) to the original images were performed to increase the total amount of data. Following training on a total of 2772 images, data augmentation techniques were applied twice, once for validation and once for testing. The CNN architecture of Mendes et al. consisted of four blocks: convolution, batch normalisation, ReLU, and max pooling. After these blocks, the network concluded with two FC layers, representing the two classes in the dataset: one with 128 units and another with 2 units. A softmax layer was then mapped to the outputs of the last layer. The Adam Optimizer was applied for the learning process, and its learning rate was set at 0.001. A 93.2% accuracy was achieved by the model on the testing set, with noteworthy results for sensitivity (92%), specificity (94%), precision (94%), F1-score (94%), and Cohen's kappa (0.86). However, their model was trained on a relatively small dataset (only 77 volumes), which limits generalizability. Moreover, their approach relies on selecting the most visible lesion slice, potentially overlooking valuable contextual information from adjacent slices.

In 2024, Du et al. (2024) addressed the class imbalance challenge in DBT, where abnormal cases occupy a small fraction of the imaging volume. The authors proposed SIFT-DBT, a method combining self-supervised contrastive learning and patch-level-MIL model to improve the classification accuracy and to maintain high spatial resolution. The DBT data were used by the self-supervised learning model to form robust pairs, treating different slices from the same volume and views as positive pairs, enabling the model to focus on structural and semantic information within the images. A local multi-patch strategy was introduced for fine-tuning to maintain high resolution and optimize the computational efficiency. Their model was evaluated on the BCS-DBT dataset, classifying images into normal and abnormal (including benign and malignant classes). They achieved an AUC of 92.69%, a specificity of 84.15%, and a sensitivity of 84.62%. Their system primarily focuses on volume-level labels without exploring fine-grained lesion annotations. Additionally, while effective in boosting performance, the approach's reliance on multiple training stages and local patch selection may hinder scalability and clinical implementation.

Farangis Sajadi Moghadam and Rashidi (2024) developed a novel feature extraction model based of Discrete Cosine-based Stockwell Transform (DCT-DOST) and radiomic features to classify DBT images into benign and malignant. Their study methodology involved four key stages: image processing, tumour segmentation, feature extraction, and classification. 713 radiomic features and 3304 DCT-DOST features were extracted from the ROI. Synthetic Minority Oversampling Technique (SMOTE) was implemented for feature selection. Classification was performed using RF, KNN, and SVM algorithms. The best results from their system were achieved when employing RF classifier, with an accuracy of 78.51%, AUC of 87.80%, sensitivity of 82.78%, and specificity of 75.19%. These results demonstrate that integrating DCT-DOST and radiomic features enhance the classification of tumours. However, their reliance on handcrafted features and classical machine learning classifiers may limit adaptability to complex patterns in DBT data. Furthermore, the dataset used was relatively small and lacked external validation, raising concerns about generalizability to broader clinical settings.

Research by Farangis et al. (2023) focuses on the classification of benign and malignant tumours in DBT images using Radiomic-based methods. Their research

leverages BCS-DBT dataset and applies advanced Radiomic feature extraction techniques to classify breast tumours. The central slice of the DBT image, containing key anatomical details, is used for feature extraction, and the study evaluates various machine learning algorithms, including Random Forest and Quadratic Discriminant Analysis (QDA). Among these, QDA achieved the best results with an AUC of 88.56%, accuracy of 88.67%, and sensitivity of 77.12%. Their system relied on handcrafted feature extraction, which may not capture the full complexity of tumour heterogeneity. Additionally, the study lacked cross-dataset validation, limiting confidence in the model's generalizability across different clinical environments.

Hassan et al. (2024) introduced a novel deep learning framework for classifying breast tumours in DBT images by combining image quality-aware features and tumour texture descriptors. Unlike most existing methods, this approach considers the image quality degradation caused by artifacts like patient movement and low radiation doses, which can affect the accuracy of automated classification systems. The framework employs a two-branch model, where one branch extracts tumour texture descriptors using a CNN, while the other, named TomoQA, focuses on assessing the overall image quality. By integrating both texture and quality-aware features, the model significantly improves classification performance, achieving an accuracy of 78.26% and a precision of 88.24%. The study shows that incorporating image quality assessment into the classification process enhances the ability of the model to distinguish between benign and malignant tumours. However, the performance of the model may be affected by limited external validation, and the reliance on quality assessment modules could introduce variability due to differences in acquisition.

4.4 Summary

The literature study explores thoroughly the classification of DBT scans, focusing on the application of deep learning models and CAD Systems. Numerous studies using state-of-the-art technology are conducted in this field of study with the goal of improving the accuracy of breast cancer classification. The chapter starts with a thorough analysis of CAD Systems in relation to DBT scans and then moves on to the development and assessment of complex systems meant to automatically classify lesions in DBT images. Next, a detailed review of the studies that used the

BCS-DBT dataset (Buda et al., 2020) is conducted. This dataset provided the data used in the development of the systems produced in this PhD study.

In the reviewed literature, the study conducted by Nogay, Akinci, and Yilmaz (2021) was the only study that tackled the challenge of multi-class classification of DBT images. Pre-trained DCNN models were used and classified the images into normal, benign, or malignant. AlexNet achieved highest accuracy of 75%, but only accuracy statistics were reported, without including additional performance measures such as specificity, sensitivity, or other relevant metrics. Without specificity and sensitivity, it is unclear how effectively their system identifies true abnormal cases (benign or malignant) or avoids misclassifying normal cases as abnormal.

The remaining studies focused on binary classification, either differentiating between normal and abnormal cases(where “abnormal” includes both benign and malignant cases) or between cancerous and non-cancerous cases (where “non-cancerous” includes both benign and normal cases), or between benign and malignant cases. While some studies achieved high accuracy rates, they lack the ability to provide a more detailed classification of DBT scans, a distinction crucial for decision making. This highlights a clear gap in the field, where multi-class classification is more clinically significant.
In the following chapter, the dataset considered in this PhD study will be introduced, providing details on the composition of the dataset, the volume of available data, and the classification criteria applied to the dataset. Furthermore, the experimental configuration will be described, providing a thorough account of the performance metrics utilised to evaluate the work and the systems developed throughout this investigation.

Chapter 5 Dataset and Experimental Framework

5.1 Introduction

The dataset, experimental setup, and performance metrics utilised within the research are introduced in this section. This first subsection provides a detailed explanation of the structure of the dataset, including a list of the cases that have been included and a description of their distribution across different classes. The subsequent subsection then elaborates on the experimental setup and clarifies the subsets that were selected for each system. Finally, the final section clarifies the performance measures employed to assess the performance of the system, providing the corresponding formulas and outlining the methodological approach for their calculation.

5.2 Dataset Description

This study utilized data from the Breast Cancer Screening - Digital Breast Tomosynthesis (BCS-DBT) collection, which is available through The Cancer Imaging Archive and was detailed by Buda et al. (2020). Although digital breast tomosynthesis is a widely researched area within AI medical imaging, the progress and testing of algorithms in this field are often limited by a shortage of large, well-annotated datasets that are publicly accessible (Buda et al., 2021; Mota et al., 2022). The BCS-DBT dataset was selected for this research due to its widespread recognition as one of the leading options for breast cancer detection and classification using tomosynthesis. The high-resolution 3D images provide a significant advantage, enabling the detection of subtle signs that might be missed by traditional 2D mammography, thus enhancing classification accuracy. As the BCS-DBT dataset continues to be adopted by an increasing number of researchers, it is becoming a standard reference in the field, aiding in the validation of new methods and contributing to advancements in breast cancer classification and detection. The DBT volumes originated specifically from Duke University Hospital and Duke University in Durham, North Carolina, USA. According to the research by Buda et al. (2020), a total of 16,802 DBT examinations performed between August 26, 2014, and January 29, 2018, were received from Duke Health System for the purposes of this diagnostic investigation. The entire dataset consists of 22,032 breast tomosynthesis scans that were collected from 5,060 patients in total. According to

Buda et al. (2020), the dataset is divided into four different case categories: normal, actionable, benign, and malignant. The three categories of normal, benign, and malignant tissue—all of which had undergone biopsy procedures—were the only subject of this research. The number of patients and scans in each partition is shown in Table 5.1 and Table 5.2.

Table 5.1 The number of patients in each category from the (BCS-DBT) dataset (Buda et al., 2020)

Sets	Normal	Benign	Malignant	Actionable
Training	4,109	62	39	178
Validation	200	20	20	40
Testing	300	30	30	60
Total	**4,609**	**112**	**89**	**278**

According to Buda et al. 2020, the Table 5.1 provides an overview of the patient distribution across several categories using DBT scans data from the BCS-DBT dataset. Four separate categories—Normal, Benign, Malignant, and Actionable—are used to group the dataset. The figures show how many patients were in each category during the testing, validation, and training stages. There were 4,388 patients in the training set, comprising 4,109 normal cases, 62 benign cases, 39 malignant cases, and 178 actionable instances. There were 200 normal cases, 20 benign cases, 20 malignant cases, and 40 actionable cases total—a total of 280 patients—in the validation set. The testing set included 420 patients in total: 300 normal cases, 30 benign cases, 30 malignant cases, and 60 actionable instances. The dataset as a whole consists of 278 actionable instances, 89 malignant cases, 112 benign cases, and 4,609 normal cases.

As stated by Buda et al. in 2020, Table 5.2 shows the distribution of scans throughout the various segments of the BCS-DBT dataset. The number of scans in each segment is shown in the table as a proportion of the entire dataset size as well as raw counts. Of all the datasets, 87% are found in the training segment, comprising 19,148 scans. 1,163 scans comprise the validation partition, or 5% of the whole dataset. The testing segment encompasses 1,721 scans, contributing 8% to

the total dataset. Summing up the training, validation and testing segments, the dataset encompasses a total of 22,032 scans. The number of slices required by breast tomosynthesis patients might vary depending on several specific criteria, most notably the distinctive features of each breast scan. This variance is caused by various factors, including overall structure, density, and size of the breasts. When breast density is higher, more image slicing could be required during the tomosynthesis scan in order to guarantee a comprehensive and in-depth analysis.

During a typical DBT exam, patients generally have two scans per breast—one from the top (craniocaudal or CC view) and one from the side (mediolateral oblique or MLO view), totalling four scans. Since each scan can produce between 40 to 100 images, a patient might end up with anywhere from 160 to 400 images from the entire exam.

Table 5.2 The number of scans in each partition from the (BCS-DBT) dataset (Buda et al., 2020).

Sets	Number of scans
Training	19,148 **(87%)**
Validation	1,163 **(5%)**
Testing	1,721 **(8%)**
Total	**22,032**

5.3 Experimental Setup

Three distinct subsets were used in the development of the models devised in this dissertation, due to the massive number of cases and the enormous dataset size (1.526 TB). The information and patient count for each subset are provided in Table 5.3. Several measures were used to compare the performance of each system on each subset. In this study, 80% of the data was used for model training and cross-validation. This 80% subset was subjected to 10-fold cross-validation. The remaining 20% of the data was held out as a separate test set, entirely isolated from the training and cross-validation process. The proposed systems were developed using

MATLAB R2021a. The run-time hardware platform was a computer fitted with a 2.8 GHz Quad-Core, Intel Core i7, 16 GB RAM, 1 TB Storage, and an Intel Iris Plus Graphics 655 (1536 MB). Several performance metrics were considered in this research, including accuracy, and the weighted average sensitivity, specificity, precision, and the run time.

An overview of the patient distribution across several subsets used in the analysis of the developed systems is shown in Table 5.3. The subsets—Normal, Benign, and Malignant—are grouped according to the findings of the diagnostic process. There are 200 patients in Subset 1 overall, of which 99 have normal findings, 62 are benign, and 39 are malignant. Subset 2 includes 199 patients with normal results, 62 cases of benign, and 39 cases of malignancy, for a total of 300 patients. Subset 3 comprises a total of 600 patients, of which 499 have normal findings, 62 are benign, and 39 are malignant.

To ensure the integrity and validity of the system, the entire case of each patient (all related scans) was assigned to a single group-either training, validation, or testing. This approach prevents any overlap of scans from the same patient across multiple groups, thereby preventing potential bias that could arise from correlated images. To ensure that the system learns general patterns rather than patient-specific features, scans from the same patient are kept in one group.

Table 5.3 *Number of patients in each data subset used in the investigation of the developed systems.*

	Subset 1	**Subset 2**	**Subset 3**
Normal	99	199	499
Benign	62	62	62
Malignant	39	39	39
Total	**200**	**300**	**600**

5.4 Performance Indicators

To evaluate the effectiveness of the developed systems, various performance metrics were taken into consideration. Evaluating the performance of the systems mainly depends on building the confusion matrix. The confusion matrix for a two-class classification model is shown in Table 5.4.

When evaluating a classification model in the context of binary classification, which entails classifying instances into one of two classes (typically represented as positive and negative), the terms True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) are utilised.

Table 5.4 Two-class confusion matrix

		Predicted	
		Positive	Negative
Actual	Positive	**True Positive**	**False Negative**
	Negative	**False Positive**	**True Negative**

- **True Positive (TP):**
  - Real positive instances that the model accurately classifies as positive.
  - Example: A genuine positive in a diagnostic imaging test for abnormalities would be an image displaying an abnormality that the diagnostic model accurately identifies as such.
- **False Positive (FP):**
  - When an instance does not show any abnormalities, but the model classifies it wrongly as having abnormalities.
  - Example: A FP in a diagnostic imaging test for abnormalities would be an image that shows no abnormalities but is mistakenly classified as having an abnormality by the diagnostic model.
- **True Negative (TN):**
  - Those instances where the model accurately classifies them as normal despite the absence of anomalies.

- Example: A TN in a diagnostic imaging test for abnormalities would be an image that the diagnostic model accurately classified as normal yet contains no abnormalities.

- **False Negative (FN):**
  - Observations that the model mistakenly classifies as normal but in fact show anomalies.
  - Example: A FN in a diagnostic imaging test for abnormalities would be an image displaying an abnormality that the diagnostic model misidentifies as normal.

An explanation of the significance of each metric is presented below (Grandini et al., 2020):

1. **Accuracy**

Considers both true positives and true negatives to assess the overall accuracy of the classification.
Significance: Offers a broad summary of the effectiveness of the model in every class.

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

2. **Sensitivity / Recall**

Calculates the percentage of real positives that are appropriately recognised.
Significance: Very important when false negatives come at a great cost.

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

3. **Precision (PPV - Positive Predictive Value):**

Shows the percentage of positive forecasts that were accurate out of all the positive predictions.
Significance: Relevant in situations where the expense of false positives is substantial.

$$Precision = \frac{TP}{TP+FP} \quad (5.3)$$

4. **Specificity**

Calculates the percentage of real negatives that are appropriately recognised.
Significance: Considerable when considering the expense of false positive results.

$$Specificity = \frac{TN}{TN+FP} \quad (5.4)$$

5. **F1-score**

Harmonic mean of precision and sensitivity, providing a balanced metric.
Significance: Useful when there is an imbalance between classes.

$$F1 - Score = \frac{2*Precision*Sensitivity}{Precision+Sensitivity} \quad (5.5)$$

The formulas that were previously presented were generated for the purpose of assessing a two-class classification model. These metrics have been calculated separately for every class in the framework of the three-class classification model. The following formulas are used to calculate the Weighted Average Precision, Recall, F1 score, and Specificity (Grandini et al., 2020) in this multi-class classification model to thoroughly evaluate the whole system:

1. **Weighted Average Sensitivity**

$$Weighted\ Avg.Sensitivity = \frac{\sum_{i=1}^{C} w_i*Sensitivity}{\sum_{i=1}^{C} w_i} \quad (5.6)$$

2. **Weighted Average Precision**

$$Weighted\ Avg.Precision = \frac{\sum_{i=1}^{C} w_i*Precision}{\sum_{i=1}^{C} w_i} \quad (5.7)$$

3. **Weighted Average Specificity**

$$Weighted\ Avg.Specificity = \frac{\sum_{i=1}^{C} w_i*Specificity}{\sum_{i=1}^{C} w_i} \quad (5.8)$$

4. **Weighted Average F1-Score**

$$Weighted\ Avg.F1 - Score = \frac{\sum_{i=1}^{C} w_i*F1-Score}{\sum_{i=1}^{C} w_i} \quad (5.9)$$

Where:

- ***C*** : The number of classes
- $\boldsymbol{w_i}$ **:** Weight of class i

These formulas incorporate weights for each class in the calculation of the weighted averages, providing a comprehensive evaluation of the multi-class classification model.

In several fields, evaluating the relative significance of improvements made to one system over another or from an earlier version is an essential component. This assessment aids in evaluating the effectiveness of system modifications and offers information on whether those changes result in significant improvements. For assessing significance, Cohen's d and the t-test were employed using multiple performance metrics, including accuracy, sensitivity, specificity, precision, and F1 score. This multidimensional evaluation approach ensures that the observed improvements reflect genuine diagnostic capability rather than just improvements in accuracy.

1. **Cohen's d**

A statistical metric named Cohen's d is employed to evaluate the impact magnitude of the distinction between two groups or scenarios. Cohen's d is a useful metric for assessing the significance of improvement between two systems. It is utilised to assess the degree to which important performance parameters such as accuracy, sensitivity, specificity, F1 score, and precision have improved. This statistical measure provides insight into the practical importance of observed improvements by providing a standard expression for the size of the difference between the mean values of the two groups. Comparing the means of the two groups and normalising the difference by the pooled standard deviation is the process of calculating Cohen's d (Nakagawa and Cuthill, 2007).

When considering within-group variability, Cohen's d measures how much the mean performance of one model differs from another on several performance metrices. A larger effect size, which denotes a more noticeable and practically significant gain in performance, is indicated by a higher Cohen's d value. The mathematical expression for Cohen's d is as follows:

$$d = \frac{\overline{X_1} - \overline{X_2}}{S_{pooled}} \qquad (5.10)$$

Where:

- $\overline{X_1}\ and\ \overline{X_2}$ are means of the two groups (Performance measures for the systems).
- S_{pooled} is the pooled standard deviation, calculated as:

$$S_{pooled} = \sqrt{\frac{(n_1-1) * s_1^2+(n_2-1).s_2^2}{n_1+n_2-2}} \qquad (5.11)$$

Where:

- n_1 and n_2 represent the sample size of the two groups.
- s_1 and s_2 are their respective standard deviations.

An example of this kind of effect size index is Cohen's term d. Effect sizes were categorised by Cohen as small (d = 0.2), medium (d = 0.5), and large (d > 0.8). A small effect size indicates slight variation with few practical implications. A large effect size denotes a considerable and practically meaningful difference between the means, whereas a medium effect size implies a perceptible but mild difference. Tomosynthesis classification models that use Cohen's d allow for a more thorough evaluation of the importance of improvements in performance measures. Beyond conventional measures of statistical significance, the measure provides a standardised metric to assess the practical relevance of modifications, enabling a comprehensive analysis of the observed differences.

2. t-test

A prominent statistical technique for determining if the mean difference between two groups or conditions is statistically significant is the t-test. When evaluating the effectiveness of two systems, it is particularly beneficial. The ratio of the difference between the means to the standard error of the difference is represented by the t-statistic, which is computed by the t-test. The following is the t-statistic formula:

$$t = \frac{\overline{X_1}-\overline{X_2}}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \qquad (5.12)$$

Where:

- $\overline{X_1}\ and\ \overline{X_2}$ are means of the two groups (Performance measures for the systems).
- n_1 and n_2 represent the sample size of the two groups.
- s_1^2 and s_2^2 are their respective variance.

The p-value, which the t-test calculated, represents the probability to observe a difference between the two groups in the event that there isn't one. When a p-value is minimal, usually less than 0.05, it indicates that the observed difference cannot be explained by chance, which means the null hypothesis—that there is no significant difference—is rejected. The degrees of freedom (*df*) and the t-statistic form the basis of the p-value formula.

$$p - value = P(T \geq |t|). \quad (5.13)$$

Where:

- T follows a t-distribution with *df* = $n_1 + n_2 - 2$.

Both the Cohen's d and the t-test are statistical techniques used for many types of data analysis, especially when determining how significant differences are between two groups. Cohen's d offers a standardised measure of effect size, clarifying the practical relevance of observed differences, whereas the t-test concentrates on evaluating if the means of two groups are significantly different. Analysing the significance of the mean difference between two groups can be done simply with the t-test. Because it is sensitive to changes in sample size and gives a precise p-value, it helps calculating the likelihood that the observed difference was the result of pure chance. The shortcomings of the t-test, however, include its sensitivity to assumptions about normality and variance homogeneity, which may compromise its validity in specific circumstances.

However, Cohen's d, which expresses the size of the observed difference in standard deviation units, offers a measure of effect size. This makes it possible to assess the practical significance of the differences in addition to its statistical significance. Cohen's d, which offers a standardised metric, is useful for

comparing with different metrics. The t-test provides the advantages of being straightforward, simple to comprehend, and able to clearly show statistical significance. It is especially beneficial if there is a substantial difference between two groups. On the other hand, Cohen's d is useful for making comparisons between various studies and metrics when the focus is on comprehending the practical significance of the observed difference.

To avoid misleading conclusions, the evaluation was not limited to accuracy alone. Improvements in accuracy were considered alongside sensitivity, specificity, precision, and F1 score to ensure a more balanced and accurate assessment of the model's performance. By evaluating multiple performance metrics, a more comprehensive understanding of the model's diagnostic capability was obtained.

Using both Cohen's d and the t-test provided a complete evaluation of model improvements. While the t-test confirmed statistical significance, Cohen's d highlighted the practical relevance of the observed improvements. This combined approach ensured that performance gains were both statistically and practically meaningful.

5.5 Summary

This chapter presented a thorough explanation of the dataset used in this dissertation research, including information on the experimental setup, subsets, along with the respective case quantities within each class. A discussion of the performance metrics used to assess the developed systems was then provided, along with an explanation of each statistic and the methodology for their calculation. Furthermore, the critical evaluation of the impact of improvements between systems, whether in comparison or across iterations, was emphasised in this chapter. This assessment was helpful in determining the efficacy of the system changes and provided information on the magnitude of the improvements that were reported. Metrics like Cohen's d and the t-test were emphasised in this chapter as essential instruments frequently used to determine the statistical significance of differences, providing useful measurements for carrying out in-depth evaluations of system performance and modifications.

The developed systems will be presented in the following chapters, which will start with an explanation and system diagram for each, followed by an in-depth description of the corresponding applied models. The full disclosure of all results across performance indicators will be provided next. Further exploration of the connections and a thorough explanation of the rationale for the development of each system will be provided in later chapters.

Chapter 6 Comparative Evaluation System for Deep-Learning Models for Feature Extraction

6.1 Introduction

In this chapter, a comparative evaluation system for deep-learning models for feature extraction is introduced. This primary focus of this system was to evaluate the performance of state-of-the-art deep learning models in terms of feature extraction from DBT images. This comprehensive system included DBT augmentation, image enhancement techniques, and colour feature mapping for tissue separation. Utilising six state-of-the-art deep learning models, namely ResNet-18, AlexNet, GoogleNet, MobileNetV2, VGG-16, and DenseNet-201, the system aimed to extract discriminative features from DBT slices. The combination of these models was strategically chosen to capitalise on their diverse architectures and strengths. Subsequently, the extracted features were employed in a SVM classifier to effectively classify DBT slices.

6.2 Methodology

As stated in Chapter 3, CAD Systems consist of sequential modules. The system introduced in this chapter includes data augmentation for images followed by images enhancement in the pre-processing stage. After enhancing the augmented images, images are input to a colour feature map stage to separate image intensities. Features are then extracted and input to a classifier for a multi-class classification of DBT images. A comparison was conducted to evaluate the effectiveness of the six deep-learning models, with Figure 6.1 illustrating the CAD System employed for the assessment.

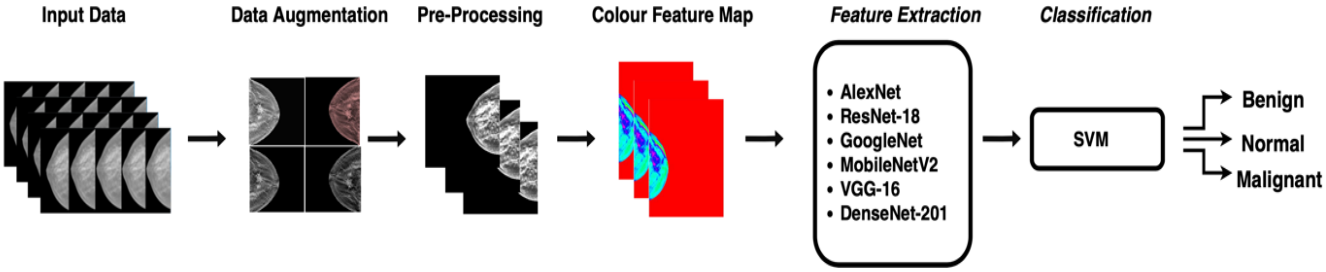

Figure 6.1 Developed DE System

Data augmentation was applied to raw input images in the first phase of this system by flipping, rotating, and applying random values for brightness, saturation, and hue. For this experiment, four images were generated from each tomosynthesis slice. Interestingly, out of the three subsets that were used, two of the images were flipped while the other two were not. A total of 41,632 slices across all classes were retrieved from Subset 1, which included 39 patients with malignant abnormalities, 62 patients with benign findings, and 99 normal cases. The first stage in getting slices ready for the ensuing training, validation, and test sets was the augmentation procedure. The total number of slices in Subset 1 after augmentation was 166,528. Similarly, Subset 2 included 65,050 slices from all classes, including 199 normal cases, 62 patients with benign findings, and 39 patients with malignant abnormalities. After augmentation, The total number of slices in Subset 2 increased to 260,200. Finally, Subset 3 consisted of 499 normal cases, 62 patients with benign findings, and 39 patients with malignant abnormalities, resulting in a cumulative total of 147,632 slices across all classes. Post-augmentation, the total number of slices in Subset 3 amounted to 590,528. The augmentation procedure was applied to all systems developed in this thesis and will be presented in the upcoming chapters.

The augmented images were enhanced during the pre-processing step following the augmentation phase. By redistributing pixel intensities, histogram equalisation for contrast enhancement was implemented to improve the visibility of subtle features in DBT images. Additionally, to reduce undesired artefacts and improve the overall quality of the images, a noise reduction technique such as Gaussian smoothing, using a Gaussian filter with a 0.5 standard deviation, was applied to reduce noise while preserving details. These subtle smoothing averages pixel values with neighbours, minimizing noise without significantly blurring edges or masses or calcifications. Samples of normal, benign, and malignant cases are shown in Figure 6.2 after applying the pre-processing techniques.

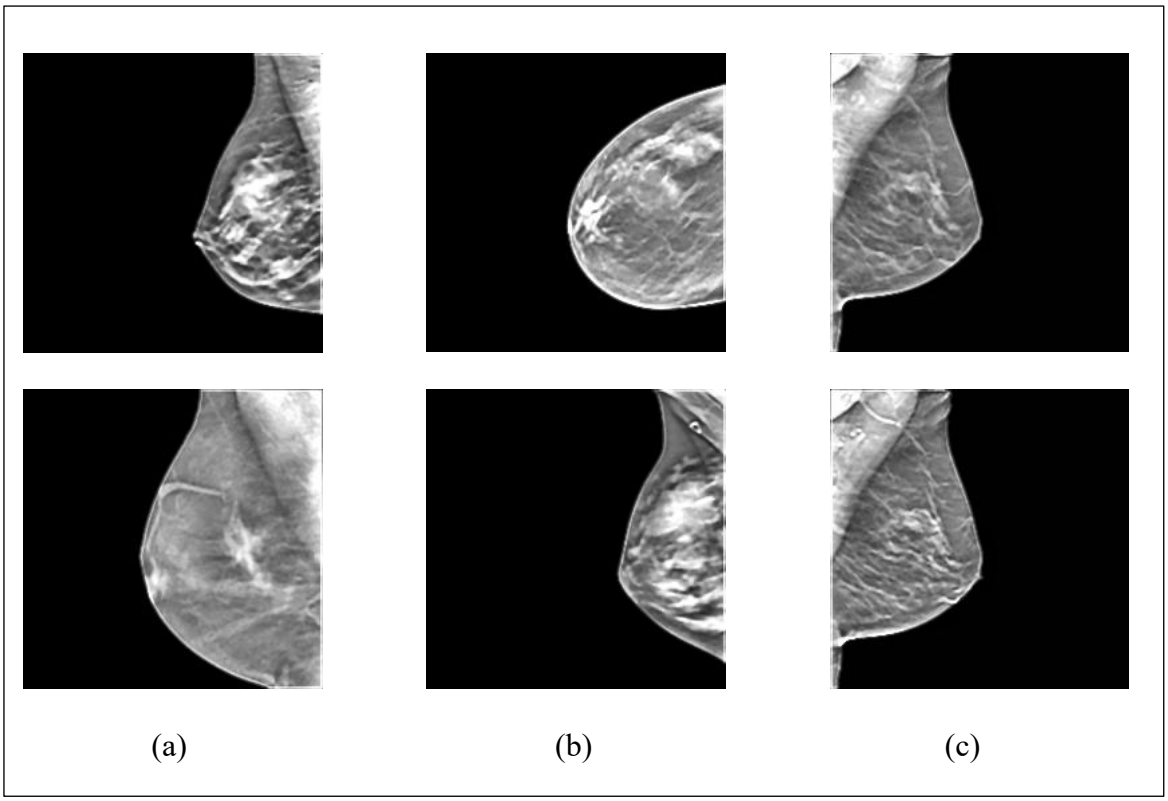

Figure 6.2 Samples of (a) Benign cases (b) Malignant cases (c) Normal cases after pre-processing

HSV colour feature map has been utilised after the pre-processing stage to aid the classifiers in better differentiating between various tissue types in the scan. HSV feature maps in this colour feature map draw attention to dominating colours, vividness, and brightness, respectively. Through the application of the HSV feature maps, the image classification model gains access to rich colour data in its input, improves the ability of the model to distinguish between various breast tissues. Samples of normal, benign, and malignant cases are shown in Figure 6.3 after applying the colour feature map.

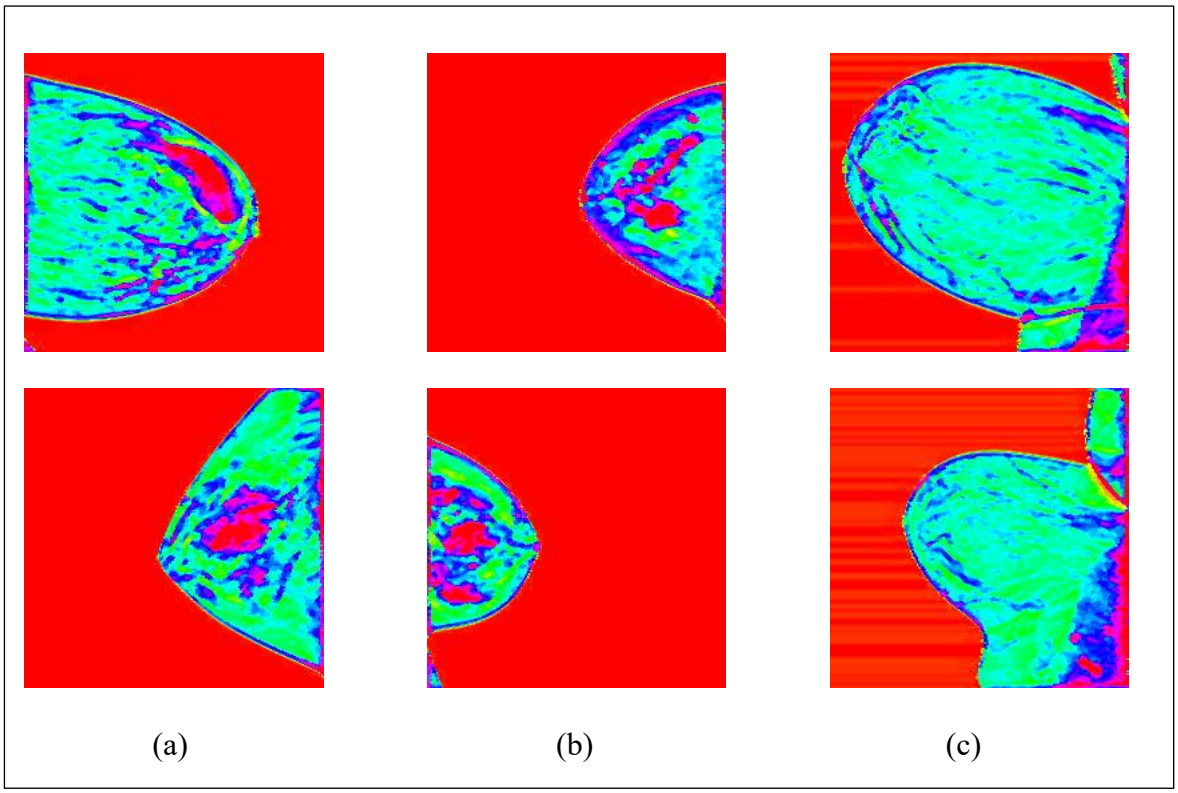

Figure 6.3 Samples of (a) Benign cases (b) Malignant cases (c) Normal cases after the colour mapping technique

Six cutting-edge deep learning models, including AlexNet, ResNet-18, GoogleNet, MobileNetV2, VGG-16, and DenseNet-201, were used to extract features after the colour map model was applied. More specifically, the final convolutional layer of each deep learning model, also known as the deep or high-level layer, provided the features. This layer is very useful for improving the accuracy of differentiating between the various classes in the tomosynthesis classification because it is proficient at capturing complex and abstract representations of the input images (Yasaka et al., 2018) (Mostafa and Wu, 2021). A more thorough representation of the entire content and context of the images is provided by the features that were extracted out of the final convolutional layer. This strong representation helps to accurately classify breast tomosynthesis images by helping to discern between the normal, benign, and malignant classes. The decreased vulnerability to noise and local variations in the final convolutional layer, helps to provide a more robust representation of the underlying structures in the breast tomosynthesis images, further supports the decision to extract features from this layer. The closeness of the final convolutional layer to the classification layer is a key element supporting this decision. Conceptually complex features that hold information crucial to the final

image classification task are extracted from this layer. By utilising these high-level features as input to subsequent FC layers or the classification models, the classification process benefits from a more discriminative and informative representation. This approach improves decision-making in the process of classifying DBT images, which in turn improves the ability of the model to correctly classify images into benign, malignant, and normal classes.

The SVM classification model is the final model that is applied in the classification process. The SVM classifier plays a crucial role in classifying DBT images into three categories: malignant, benign, and normal. SVMs are ideally suited for this application due to their ability to handle multi-class classification challenges and learn complicated patterns within high-dimensional feature spaces (Foody and Mathur, 2004). The advantages of using SVM for DBT image classification include robust performance, even when there is a lack of training data, and the ability to manage non-linear connections in the features of the image (Guido et al., 2024). The SVM classifies the features, ultimately resulting in the final classification of normal, benign, and malignant classes. Chapter 3 provides a comprehensive explanation of all the models employed by this system. Results are presented and discussed in the following subsection.

6.3 Results and Discussion

Evaluating the effectiveness of state-of-the-art deep learning models for feature extraction from DBT images was the primary objective of the first developed system. This all-inclusive strategy comprised colour feature mapping, DBT augmentation, and image enhancing techniques to enhance tissue classification. The approach aims to extract distinctive features from DBT slices by incorporating six well-known deep learning models: ResNet-18, AlexNet, GoogleNet, MobileNetV2, VGG-16, and DenseNet-201. These models have been meticulously considered, making use of their distinctive features and varied structures. The extracted features are utilised as input into a SVM classifier, to effectively categorise DBT slices after feature extraction. Metrics including accuracy with 95% confidence intervals (CIs), sensitivity, specificity, F1-score, and precision were used to assess the performance of the system and provide an in-depth understanding of how beneficial it is in the classification of DBT scans.

Three separate data subsets were generated and utilised in the developed systems, as described in the Experimental Setup section of Chapter 5. The results of DE System utilising these three subsets are shown in Tables 6.1, 6.2, and 6.3.

An in-depth analysis of the performance of the DE System, as presented in Table 6.1, reveals unique patterns among the applied deep learning models when evaluated on Subset 1. AlexNet is the best performer, with the highest F1-Score 55.66% and accuracy 56.86% (±6.86%). As evidenced by its balanced sensitivity 56.86% and specificity 72.34%, this model demonstrates an excellent balance between precision and recall. The high specificity reflects proficiency of AlexNet in reliably recognising normal cases. On the other hand, DenseNet-201 exhibits the lowest F1-Score 46.59% and accuracy 46.11% (±6.91%), suggesting possible limits in its capacity to correctly classify cases within Subset 1. Effective capture of genuine positives is hampered by the lower sensitivity 46.11% and higher misclassification rate 53.89%. Despite being competitive at 68.95%, specificity of this model cannot make up for its overall inferior accuracy. ResNet-18, GoogleNet, VGG-16, and MobileNetV2 fall within a moderate performance range, with accuracies ranging from 47.60% (±6.92%) to 51.53% (±6.93%). Although VGG-16 has a slightly higher specificity 70.28%, the competitive F1-Score of 50.71% of MobileNetV2 can be attributed to its balanced sensitivity and accuracy. All performance metrics show comparable results between ResNet-18 and GoogleNet.

Table 6.1 The performance of CAD DE System using different deep learning models for feature extraction, assessed on Subset 1

Deep Learning Model	Performance Measures for DE System on Subset 1				
	Accuracy	**Sensitivity**	**Precision**	**Specificity**	**F1-Score**
ResNet-18	48.65%	48.65%	48.32%	68.72%	48.45%
AlexNet	56.86%	56.86%	55.21%	72.34%	55.66%
GoogleNet	47.60%	47.60%	47.46%	68.05%	47.41%
VGG-16	50.67%	50.67%	50.68%	70.28%	50.67%
MobileNetV2	51.53%	51.53%	50.11%	68.75%	50.71%
DenseNet-201	46.11%	46.11%	47.46%	68.95%	46.59%

In conclusion, complex insights of this comparison emphasise how important it is to consider a variety of indicators to fully evaluate model performance. Applied to Subset 1, AlexNet is the best-performing deep learning model, indicating its potential applicability for breast cancer classification in CAD DE System. On the other hand, DenseNet-201 could need to undergo careful evaluation and possible improvement to improve its precision and diagnostic usefulness.

Following an analysis of the performance of the DE System on Subset 2 as given in Table 6.2, AlexNet achieves a maximum accuracy of 77.80% (±4.70%) in differentiating itself, indicating its exceptional ability to precisely identify occurrences in Subset 2. Additionally, it obtains a remarkable F1-score of 75.46%, indicating a decent trade-off between recall and precision. Nonetheless, ResNet-18 and VGG-16 show competitive results with 67.43% (±5.30%) and 68.89% (±5.24%) percent accuracy, respectively. In comparison to ResNet-18, VGG-16 exhibits better specificity and precision, demonstrating its ability to reduce FPs. Accuracy values in the low 60s have been demonstrated by both GoogleNet and MobileNetV2, with GoogleNet marginally beating MobileNetV2. Despite having an average accuracy of 65.40%, DenseNet-201 stands out among the models with the highest specificity 51.80%, indicating that it is effective at correctly identifying negative cases.

Table 6.2 The performance of CAD DE System using different deep learning models for feature extraction, assessed on Subset 2

Deep Learning Model	Performance Measures for DE System on Subset 2				
	Accuracy	**Sensitivity/Recall**	**Precision**	**Specificity**	**F1-Score**
ResNet-18	67.43%	67.43%	70.28%	44.20%	68.76%
AlexNet	77.80%	77.80%	73.44%	43.55%	75.46%
GoogleNet	63.08%	63.08%	69.18%	44.53%	65.61%
VGG-16	68.89%	68.89%	70.53%	43.91%	69.63%
MobileNetV2	68.42%	68.42%	70.02%	41.32%	69.17%
DenseNet-201	65.40%	65.40%	71.94%	51.80%	68.18%

In assessing model performance, the comparison emphasises how crucial it is to consider a variety of measures. Accuracy is an important metric, but in medical diagnostic settings, specificity and the balance between precision and recall are

crucial. The unique goals of the CAD System and specifications, such as the need to reduce FPs or negatives, should be taken into consideration while selecting a deep learning model. A detailed understanding of the advantages and disadvantages of each model is made possible by this thorough study, which aids in making well-informed decisions on the implementation of CAD DE System.

Table 6.3 The performance of CAD DE System using different deep learning models for feature extraction, assessed on Subset 3

Deep Learning Model	Performance Measures for DE System on Subset 3				
	Accuracy	**Sensitivity/Recall**	**Precision**	**Specificity**	**F1-Score**
ResNet-18	87.91%	87.91%	86.47%	18.88%	87.14%
AlexNet	89.60%	89.60%	87.17%	21.43%	88.34%
GoogleNet	83.47%	83.47%	86.14%	20.86%	84.76%
VGG-16	88.00%	88.00%	86.72%	20.97%	87.35%
MobileNetV2	84.47%	84.47%	86.30%	19.68%	85.36%
DenseNet-201	84.81%	84.81%	86.62%	20.68%	85.69%

Table 6.3 provides an in-depth assessment of the performance of CAD DE System on Subset 3. Compared to Subset 2, this Subset presents a distinct set of difficulties and traits, providing important information about the resilience of the models. With a high accuracy of 89.60% (±2.44%), AlexNet outperforms all other models, making it the undisputed top performer. With an F1-score of 88.34%, the model demonstrates its capacity to keep recall and precision in check. Both ResNet-18 and VGG-16 also exhibit excellent outcomes, with accuracies of 87.91% (±2.61%) and 88.00% (±2.60%), respectively. These models demonstrate a high degree of precision, demonstrating their ability to reduce FPs. Nonetheless, the specificity values of these models are rather low, indicating difficulties in accurately identifying negative situations. Conversely, GoogleNet and MobileNetV2 demonstrate consistent performance across many subsets, with accuracies in the mid-80s. DenseNet-201 exhibits balanced precision and recall, with an accuracy of 84.81% (±2.87%), as demonstrated by its 85.69% F1-score. Although higher than some models, the specificity value of 20.68% suggests that there is still space for growth in terms of accurately identifying negatives.

The comparison across Subset 3 confirms how crucial it is to test models on a variety of datasets to gauge how well they generalise. High accuracy achieved by AlexNet indicates that it can adjust to different features. The observed trade-offs between specificity and sensitivity emphasise how crucial it is to consider both the clinical setting and the unique requirements of the CAD System when choosing a model. This in-depth examination of Subset 3 performance indicators offers developers and physicians insightful information to help them decide how best to implement CAD DE System.

The differences in how the deep learning models performed come down to their design and how well they handle the specific details of DBT images. AlexNet stood out because of its relatively simple structure and how effectively it extracts important features from the images. Its use of ReLU activation helps the model learn faster, and the overlapping max-pooling improves how well it retains spatial details, both of which make it particularly good at spotting subtle patterns in DBT slices. Another reason AlexNet works so well is that it has fewer parameters than deeper models, which lowers the risk of overfitting.

DenseNet-201, on the other hand, struggled more, which makes sense considering its more complex design. Its dense connectivity, while theoretically helpful for passing information through the network, seems to create too much redundancy and make the model more sensitive to noise. This complexity likely causes overfitting, especially with smaller, more specialized datasets like DBT. ResNet-18 delivered more balanced results because its residual connections help solve the problem of vanishing gradients, but its shallow structure may limit how well it can pick up on more detailed patterns in breast tissue. Models like GoogleNet and MobileNetV2, which are designed to be more complex and flexible, performed consistently but didn’t stand out, suggesting that deeper and more sophisticated models do not necessarily outperform simpler ones when working with DBT images.

6.4 Summary

In this chapter, the primary objective was to evaluate the effectiveness of state-of-the-art deep learning models for feature extraction from DBT images. To improve tissue classification, a complete approach was applied by the developed systems, which included colour feature mapping, DBT augmentation, and image enhancing

techniques. Metrics including accuracy, sensitivity, specificity, F1-score, and precision were employed to evaluate the performance of six top-performing deep learning models: ResNet-18, AlexNet, GoogleNet, MobileNetV2, VGG-16, and DenseNet-201. These models were integrated for feature extraction and classified using SVM classifier. The system generates and uses three subsets, the results of which have been shown in Tables 1, 2, and 3. AlexNet was the top performer in Subset 1, with the highest accuracy and F1-Score. DenseNet-201 demonstrated inefficiencies indicating a need for more consideration and possible development. The results of Subset 2 demonstrated that AlexNet has remarkable accuracy and F1-score. Subset 3 demonstrated that AlexNet outperforms the other models across all performance metrics, while ResNet-18 and VGG-16 provided great precision. Specificity values, however, point to difficulties in precisely recognising both benign and malignant cases for these models. The comparison exhibited the significance of considering a variety of metrics in medical diagnostic settings and the necessity of matching model selection to the objectives and specifications of the CAD System. The results provided useful information for decision-making when CAD DE System was implemented, considering variables like minimising FPs or negatives, and adjusting to various datasets.

Specificity values across the three subsets in DE System highlighted challenges in accurately identifying benign and malignant cases, prompting consideration for the development of the next system. Furthermore, the comparative system demonstrated that AlexNet outperformed the other five state-of-the-art deep learning models in extracting informative features that more effectively discriminated between classes.

The subsequent chapter introduces the second developed system, MA System, which includes the introduction of Mod_AlexNet. Mod_AlexNet aimed to enhance performance of AlexNet in terms of accuracy and specificity, while also exploring improvements for better detection of abnormal cases. Mod_AlexNet is developed and compared with the traditional AlexNet to assess performance differences and evaluate the enhancements implemented into the modified model.

Chapter 7 Mod_AlexNet for Enhanced Diagnosis of Digital Breast Tomosynthesis

7.1 Introduction

In the preceding chapter, state-of-the-art deep learning models were examined. The investigation revealed noteworthy findings, indicating that AlexNet outperformed its counterparts, consistently achieving superior performance across all dataset subsets. Furthermore, the findings revealed a relationship between higher overall accuracy and an increase in the number of cases in the training set. But along with this increase in accuracy came an overall reduction in specificity and abnormality detection capabilities.

The novelty of this chapter lies in the introduction of a modified deep learning model, referred to as Mod_AlexNet, designed to enhance detection accuracy and elevate the classification of abnormal cases, to be able to better discriminate between benign and malignant cases.

7.2 Methodology

In this system, to ensure an equitable comparison, the augmentation, pre-processing, and colour mapping models were incorporated using identical techniques as those applied in the previous system discussed in Chapter 6. However, the deep learning model was modified. Figure 7.1 shows the CAD System, Mod_AlexNet System (MA System), used for the assessment. Initially, images underwent augmentation and enhancement, and after the enhancement, colour mapping was implemented to enhance the differentiation between various findings in the DBT scans.

The first contribution performed during this research was to modify the traditional AlexNet model, resulting in the development of Mod_AlexNet, marking a seminal contribution to the field. The aim of this architectural development was to maximise the conventional classification performance of AlexNet. Figure 7.1 highlights the new layers that have been added to develop the Mod_AlexNet. There are eight learnable layers in the AlexNet. ReLu activation is used in each of the five levels of the model, except for the output layer, which uses max pooling followed by three FC layers. Six

additional layers were added by Mod AlexNet, including 2 max-pooling layers and 4 batch normalisation levels. According to Figure 7.1, the max-pooling layers were added to the third and fourth convolution layers, after the batch normalisation layers were added after the first four convolution layers.

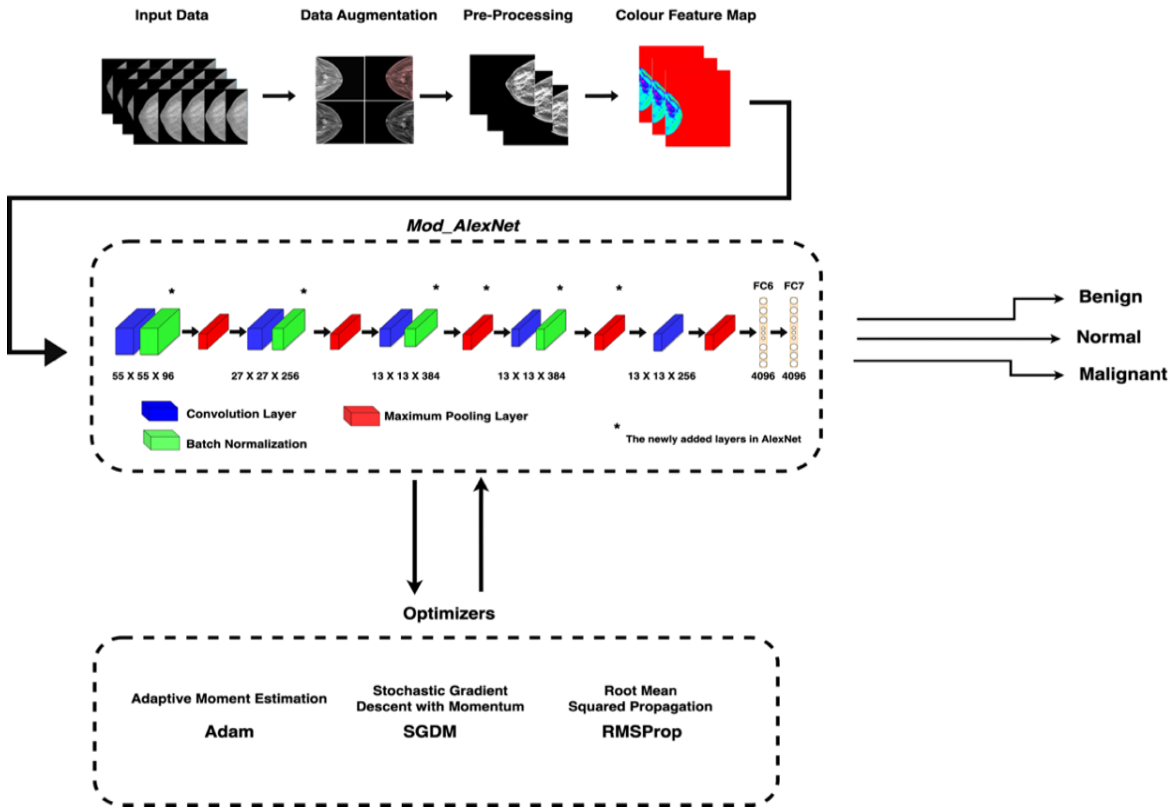

Figure 7.1 Mod_AlexNet System (MA System)

Additional layers of Mod_AlexNet are positioned strategically, as shown in Figure 7.2, to maximise classification performance, as supported by their individual benefits. By reducing internal covariate shift, the addition of batch normalisation layers following the first four convolution layers promotes a more stable training procedure (Awais, Iqbal and Bae, 2020). By preserving a steady distribution of inputs, this normalisation promotes adaptability in the model and speeds up convergence during training (Awais, Iqbal and Bae, 2020). Moreover, the feature maps are down sampled by adding max-pooling layers after the third and fourth convolution layers, which encourages spatial hierarchy and abstraction. By removing unnecessary data and capturing important properties, this down sampling technique helps the model become more abstract and less prone to overfitting (Zafar et al., 2022). The

justification for these modifications is based on a thorough comprehension of the difficulties that deep learning models provide, and the effective resolution of these difficulties depends on the thoughtful arrangement of these layers.

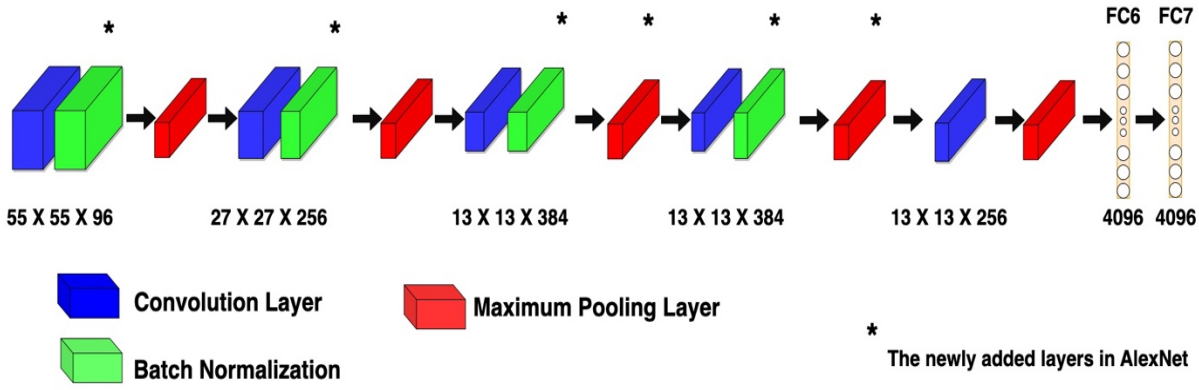

Figure 7.2 Mod_AlexNet Architecture

The ability of the network to identify spatial hierarchies in the input data is greatly improved by the addition of max-pooling layers after the third and fourth convolutional stages. By positioning these layers at this point, the model can gradually pick up on and understand complex patterns, which will help to produce a more discriminative and subtle feature representation. This hierarchical feature extraction helps the network recognise progressively complex structures in the data, which improves classification accuracy. Furthermore, the down sampling impact of max-pooling promotes a more robust and generalised model by reducing the likelihood of overfitting.

Internal covariate shift is tackled concurrently by integrating batch normalisation layers after the first four convolutional layers. At this point, normalising the inputs encourages a steady distribution of features across the network, which makes learning reliable and effective. This holds special significance for networks such as AlexNet, since it becomes more difficult to maintain stable training dynamics as the network gets deeper. By facilitating the efficient acquisition of discriminative features, faster convergence of the batch normalisation improves training efficiency and adds to the specificity of the model. The optimisation of training dynamics is strongly associated to the specific decision of positioning batch normalisation layers after the

first four convolutional layers. Early normalisation creates a solid basis for feature learning by guaranteeing that inputs to later layers have a consistent distribution. By placing the object strategically, problems like vanishing or exploding gradients are lessened, allowing information to move across layers more successfully. This makes the network less sensitive to changes in the input data, which enhances specificity and improves generalisation in classification tasks.

Within this system, an in-depth examination was conducted between the pre-trained AlexNet and its modified version, Mod_AlexNet. The main goal of this study was to determine how well these models classified enhanced images into classes that were benign, malignant, and normal. A detailed examination of model performance under various training optimisers and batch sizes was required for the comparative evaluation. This comprehensive approach was developed to identify and validate the higher classification powers that are inherent to each system. Adam, Stochastic Gradient Descent with Momentum (SGDM), and Root Mean Square Propagation (RMSProp) were the optimisers that were considered for this assessment. A variety of batch sizes were utilised for performance testing to fully evaluate the robustness of the models. The subsequent section provides an in-depth presentation of results along with a thorough discussion that examines the findings and evaluates the influence of varying optimiser configurations in conjunction with batch size.

7.3 Results and Discussion

To evaluate and compare the efficacy of the developed Mod_AlexNet with the traditional pre-trained AlexNet, a thorough examination involving multiple performance measures was carried out. The assessment included metrics like the F1 score, sensitivity, specificity, precision, and accuracy along with the corresponding CIs. To assess the statistical significance of improvement, Cohen's d was computed, along with a t-test, to compare Mod_AlexNet to pre-trained AlexNet. Additionally, the investigation involved a comparison of the output of Mod_AlexNet in the MA System to the optimal outcomes obtained from the DE System, which is presented in Chapter 6, aiming to compute the significance of improvement.

Table 7.1 Performance Measures for the implementation of AlexNet using Subset 3 in MA System

Optimizer	Batch Size	Accuracy	Sensitivity/Recall	Precision	Specificity	F1-Score
SGDM	32	88.74%	88.74%	86.78%	17.40%	87.69%
	64	88.53%	88.53%	87.42%	19.87%	87.96%
	128	89.41%	89.41%	86.03%	19.28%	87.54%
	256	89.79%	89.79%	86.70%	18.76%	88.06%
	512	90.32%	90.32%	87.59%	18.42%	88.80%
Adam	32	92.26%	92.26%	NA	7.74%	88.55%
	64	92.26%	92.26%	NA	7.74%	88.55%
	128	92.26%	92.26%	NA	7.74%	88.55%
	256	92.26%	92.26%	NA	7.74%	88.55%
	512	92.26%	92.26%	NA	7.74%	88.55%
RMSProp	32	92.26%	92.26%	NA	7.74%	88.55%
	64	92.13%	92.13%	NA	7.73%	88.48%
	128	92.26%	92.26%	NA	7.74%	88.55%
	256	92.26%	92.26%	NA	7.74%	88.55%
	512	92.26%	92.26%	NA	7.74%	88.55%

Table 7.1 illustrates the performance of AlexNet using Subset 3 in MA System under the SGDM optimiser with varied batch sizes which provides an improved comprehension of the effectiveness of the model in terms of several metrics. Surprisingly the accuracy increases steadily, peaking at 90.32% for a batch size of 512. When it comes to sensitivity or recall, the model shows an increase as batch sizes increase, going from 88.74% to 90.32%. This implies a higher degree of precision in detecting instances that are positive, which is suggestive of a positive relationship between batch size and the capacity of the model to identify real positives. On the other hand, precision, which measures the accuracy of positive predictions, slightly decreases as batch sizes increase, from 86.03% to 87.59%, indicating an implied compromise between sensitivity and precision. Specificity, a crucial indicator of how well the model detects negative occurrences, shows a complex pattern at various batch sizes. The values show a little drop with increasing batch size, ranging from 17.40% to 18.42%. This implies a possible trade-off; wherein higher batch sizes may result in a little rise in FNs. The inverse relationship between specificity and sensitivity emphasises how important it is to carefully assess the ability of the model to categorise negative cases accurately, especially in situations where reducing FNs is crucial. Simultaneously, the F1-Score, which offers a thorough evaluation of the balance between sensitivity and precision, shows a general improvement from 87.69% to 88.80%. This increasing trend implies that the

model finds a finely tuned balance between precisely detecting positive cases and reducing FPs and FNs as batch sizes increase. The growing F1-Score suggests that the general robustness of the model increases with higher batch sizes, even in the face of any potential limitations in specificity.

Moreover, Table 7.1 offers a thorough summary of the performance metrics for the use of Subset 3 in MA System for the deployment of AlexNet with the Adam and RMSProp optimisers, with different batch sizes. Interestingly, the accuracy statistic shows stable and high levels of accuracy, staying at 92.26% for all batch sizes. The sensitivity/recall values are stable at 92.26%, demonstrating the strong capacity of the system to detect positive cases. This implies that the Adam optimiser maintains a constant sensitivity/recall performance independent of batch size. Concurrently, precision values stay at 0.00% for all batch sizes, which means all cases were classified as Normal. This highlights a possible shortcoming of the model in accurately detecting abnormal cases within its predictions.

In conclusion, the findings imply that although Adam and RMSProp optimisers continuously produce excellent levels of sensitivity and accuracy, there is no precision for all batch sizes. The model may not be able to balance out the reduction of FPs and FNs, as evidenced by the low specificity and constant F1-Score. These insights are essential for improving the performance of the system and identifying areas that need to be optimised, especially for increasing specificity and precision to produce predictions that are more reliable.

Table 7.2 describes the performance metrics for Mod_AlexNet with Subset 3 in MA System, emphasising the use of various batch sizes for the SGDM optimiser. Key performance indicators such as accuracy, sensitivity/recall, precision, specificity, and F1-Score are included in the evaluation to provide insight into the performance utilising different optimisers on different batch sizes. The accuracy percentages, which range from 90.48% to 91.61%, show a steady and excellent performance.

The accuracy estimations are highly confident due to the low CIs (2.22% to 2.35%), which highlight the ability of the system to consistently provide accurate predictions across a range of batch sizes. This precision constancy is an indication of the reliability of the system. Consistently high, sensitivity ranges from 90.48% to 91.61%.

This indicates that the model is sensitive to the existence of positive examples in the dataset and consistently intercepts TPs across different batch sizes. Precision, which gauges how well positive predictions turn out, shows stability with values between 87.71% and 88.16%. Although there is a small fluctuation, the model constantly reduces FPs, which helps to support strong precision of Mod_AlexNet. Specificity reveals values in the range of 23.91% to 21.12%. This implies a moderately consistent ability to classify negative examples appropriately. The F1-Score, which balances sensitivity and precision, stays constant from 88.96% to 89.57%. This suggests a stable balance between minimising FPs and accurately detecting positive events. The ability of the system to maintain balance between sensitivity and precision is stable for a range of batch sizes.

Table 7.2 Performance Measures for the implementation of Mod_AlexNet using Subset 3 in MA System

Optimizer	Batch Size	Accuracy	Sensitivity/Recall	Precision	Specificity	F1-Score
SGDM	32	91.61%	91.61%	88.16%	23.91%	89.57%
	64	90.63%	90.63%	88.08%	21.12%	89.18%
	128	90.50%	90.50%	87.78%	22.09%	88.99%
	256	90.48%	90.48%	87.74%	22.09%	88.96%
	512	90.70%	90.70%	87.71%	22.10%	89.07%
Adam	32	92.26%	92.26%	85.12%	7.74%	88.55%
	64	89.45%	89.45%	86.87%	17.03%	88.05%
	128	90.16%	90.16%	87.27%	22.63%	88.57%
	256	90.23%	90.23%	87.18%	23.84%	88.67%
	512	90.36%	90.36%	87.29%	23.84%	88.79%
RMSProp	32	87.28%	87.28%	86.36%	20.52%	86.81%
	64	90.32%	90.32%	86.68%	17.03%	88.39%
	128	91.23%	91.23%	86.05%	17.90%	88.29%
	256	91.48%	91.48%	86.43%	18.02%	88.65%
	512	91.60%	91.60%	86.22%	14.56%	88.69%

The performance analysis of Mod_AlexNet using Subset 3 in MA System under the Adam optimiser with different batch sizes reveals complex dynamics in several important measures. The accuracy exhibits significant fluctuation, with the best accuracy recorded at a batch size of 32. The accuracy ranges from 89.45% to 92.26%. The accompanying uncertainty is highlighted by the CIs (2.14% to 2.46%), which highlight how sensitive the system is to batch size variations. The sensitivity/recall, ranging from 89.45% to 92.26%, is consistently high, demonstrating its ability to accurately detect positive events. At a batch size of 32, the highest sensitivity and maximum accuracy coincide, suggesting a strong performance in

identifying genuine positive cases but no predictions for abnormal cases. The range of precision is from 85.12% to 87.29%. Although steady, a minor decrease at bigger batch sizes points to a more complex conflict as the system seeks to minimise FPs while balancing precision. Specificity steadily rises with increasing batch sizes, from 7.74% to 23.84%. This suggests that the ability of the system to accurately categorise negative cases has improved, continuing the upward trend in precision that has been noted. In terms of accuracy and sensitivity, the F1-Score remains very stable between 88.05% and 88.79%.

Employing the RMSProp optimisers, the accuracy shows a noteworthy upward trend, rising from 87.28% at a batch size of 32 to 91.60% with a batch size of 512. This steady improvement suggests that higher batch sizes have a beneficial effect on the overall prediction accuracy. The specificity, shows a little decline with increasing batch size, going from 20.52% to 14.56%. The compromise between sensitivity and specificity is highlighted by the preference for capturing TPs, which suggests a little rise in FNs. Similar to precision, the F1-Score measure balances sensitivity and precision; it ranges from 86.81% to 88.69% and improves with increasing batch sizes.

In MA System, AlexNet performed less effectively when using the Adam and RMSProp optimisers; abnormality classes were not identified, and all cases were classified as normal. Alternatively, AlexNet achieved optimal performance with an accuracy of 90.32%, but with a specificity of 18.42% at a batch size of 512, or an accuracy of 89.41% with a specificity of 19.28%. Mod_AlexNet, on the other hand, performed more effectively, especially when applying the SGDM optimiser with a batch size of 32. It achieved the highest specificity of 23.91%, accuracy of 91.61%, and precision of 88.16%. These results validate the claim that Mod_AlexNet surpasses the pre-trained AlexNet in terms of accuracy, precision, specificity, and F1-score.

Tables 7.1 and 7.2 present the outcomes derived from the implementation of AlexNet and Mod_AlexNet using Subset 3 within MA System. A comprehensive assessment of the enhancements achieved by both the modified system and Mod_AlexNet involved the computation of statistical measures, namely Cohen’s d and the t-test. These analyses were employed to determine the significance between the optimal

outcomes in DE System and the highest-performing result in MA System, which was observed during the deployment of Mod_AlexNet optimised using SGDM with a batch size of 32. Furthermore, a comparative evaluation was conducted for all results obtained in MA System, emphasising the superior performance of Mod_AlexNet optimised using SGDM with a batch size of 32. These analyses are shown in Table 7.3.

Table 7.3 provides a comprehensive study using T-test and Cohen's d significance measures to compare the results of different models in MA System and the performance of AlexNet in DE System to assess the optimal performance of Mod_AlexNet in MA System. The statistical evaluation of model performance using Cohen's d and t-tests was conducted based on the methodology outlined in Chapter 5. This approach ensures consistency in performance evaluation across all systems. Cohen's d was used to measure the practical significance of observed differences, while t-tests were employed to determine the statistical significance of the results. The use of multiple performance metrics, including accuracy, sensitivity, specificity, and F1-score, provides a balanced evaluation of the model's effectiveness and helps to avoid misleading conclusions that might arise from focusing on a single metric. This multidimensional evaluation approach ensures that improvements in one metric do not come at the expense of another critical diagnostic factor.

Performance variances that are hardly noticeable are revealed by Cohen's d, a measure of effect magnitude. For example, the impact sizes are small (0.06127 to 0.10120) when employing AlexNet with SGDM in MA System, indicating limited practical significance. Nonetheless, the T-test significance values show that the observed improvements are statistically significant, except for a batch size of 512.

In contrast, large Cohen's d values (0.52267) of MA System suggest significant variations when using the Adam and RMSProp optimisers. Nevertheless, the T-test significance values (0.29129) imply that these variations are not statistically significant, suggesting that performance of AlexNet using these optimisers is comparable. Promoting the interpretation of Cohen's d becomes necessary when faced with this situation where Cohen's d shows a significant effect size, yet the T-test is unable to demonstrate statistical significance. Relying simply on the T-test may result in the overlooking of potentially relevant differences in such

circumstances, when a large effect size suggests practical significance, but the sample size may not be sufficient for the T-test to identify statistical significance.

Table 7.3 *Comparative Statistical Analysis using Cohen's d and T-test Significance for Evaluating Mod_AlexNet Optimal Performance in MA System against the Best Performance in DE System and All Results Obtained by Various Models in MA System.*

System/Model				Cohen's d:		T-test (p)	
System	**Model**	**Optimizer**	**Batch Size**	**Measure**	**Significance**	**Measure**	**Significance**
MA System	**AlexNet**	**SGDM**	**32**	0.10120	Small	0.02610	Significant
			64	0.08335	Small	0.01302	Significant
			128	0.08718	Small	0.00614	Significant
			256	0.07715	Small	0.02875	Significant
			512	0.06127	Small	0.10833	Not Significant
		Adam	**32**	0.52267	Large	0.29129	Not Significant
			64	0.52267	Large	0.29129	Not Significant
			128	0.52267	Large	0.29129	Not Significant
			256	0.52267	Large	0.29129	Not Significant
			512	0.52267	Large	0.29129	Not Significant
		RMSProp	**32**	0.52267	Large	0.29129	Not Significant
			64	0.52484	Large	0.28944	Not Significant
			128	0.52267	Large	0.29129	Not Significant
			256	0.52267	Large	0.29129	Not Significant
			512	0.52267	Large	0.29129	Not Significant
	Mod_AlexNet	**SGDM**	**64**	0.03459	Small	0.09038	Not Significant
			128	0.03342	Small	0.01629	Significant
			256	0.03416	Small	0.01385	Significant
			512	0.03059	Small	0.01984	Significant
		Adam	**32**	0.11340	Small	0.29828	Not Significant
			64	0.09083	Small	0.05355	Not Significant
			128	0.04086	Small	0.00046	Significant
			256	0.03199	Small	0.01706	Significant
			512	0.02863	Small	0.01744	Significant
		RMSProp	**32**	0.11185	Small	0.00234	Significant
			64	0.07827	Small	0.09539	Not Significant
			128	0.06587	Small	0.12382	Not Significant
			256	0.05697	Small	0.17667	Not Significant
			512	0.07688	Small	0.23913	Not Significant
DE System		**AlexNet**		0.058337	Small	0.00318.	Significant

The emphasis on Cohen's d in these instances is warranted because it quantifies the magnitude of observed differences, providing insight into the practical relevance of the findings. A large Cohen's d suggests that the observed effect is significant in actual terms even though statistical significance was not reached. This emphasises how crucial it is to take the larger picture into account and realise that the practical impact of an effect may still exist even in the absence of statistical significance.

Using different optimisers and batch sizes, the analysis is extended to Mod_AlexNet in MA System. Although T-test significance scores show both significant and non-significant differences, Cohen's d values are notably typically small. Mod_AlexNet, for instance, shows a notable improvement with the Adam optimiser and a batch size of 128. This highlights the subtle effects of the optimiser and batch size selections. Using the SGDM optimiser and focusing on Mod_AlexNet, a dependable pattern appears. Cohen's d values are consistently small, ranging from 0.03059 to 0.03459, across various batch sizes (64, 128, 256, and 512), indicating subtle effect sizes. The T-test significance results, however, differ. Significantly, the p-values for batch sizes 128 through 512 are 0.01629, 0.01385, and 0.01984, respectively, indicating statistical significance. This indicates that Mod_AlexNet with SGDM shows slight but significant performance gains, particularly at higher batch sizes.

When switching to Adam optimiser, Cohen's d values show moderate impact sizes since they are consistently small across batch sizes. The T-test significant values, however, present a more complex scene. With a remarkably modest p-value of 0.00046, which indicates high statistical significance, a batch size of 128 stands out. This suggests that Mod_AlexNet attains a performance level that differs noticeably from the baseline when using Adam and a batch size of 128. Other batch sizes, on the other hand, do not exhibit any statistical significance, highlighting the sensitivity of the results to certain structures.

A similar tendency is revealed by the examination of Mod_AlexNet with RMSProp in MA System. Whereas the T-test significance values show significance for smaller batch sizes (32) but not for larger ones (64, 128, 256, 512), Cohen's d values are constantly small, showing minor effect sizes. This emphasises how crucial it is to consider how batch size and optimizer interact to affect the performance variations that are observed.

Finally, to compare the benchmark performance in DE System—more precisely, using AlexNet—against the ideal performance of Mod_AlexNet in MA System. A slight but noticeable variation in performance is suggested by the Cohen's d value of 0.05834, which shows a small effect size. Concurrently, a statistically significant T-test significance value of 0.00318 is indicated. This suggests that the observed performance gap between Mod_AlexNet in MA System and top-performing model in the DE System, AlexNet, is not only statistically significant but also practically significant.

Table 7.4 Comparative Analysis of MA System Mod_AlexNet using an SGDM Optimizer (batch size: 32) and DE System AlexNet using a Variety of Metrics

System	Model	Accuracy	Sensitivity/Recall	Precision	Specificity	F1-Score
MA System	**Mod_AlexNet**	91.61%	91.61%	88.16%	23.91%	89.57%
DE System	**AlexNet**	89.60%	89.60%	87.17%	21.43%	88.34%
Percentage of improvement		2.24%	2.24%	1.14%	11.59%	1.39%

An analysis of the performance of Mod_AlexNet in MA System compared to AlexNet, in DE System, using several metrics, such as F1-Score, accuracy, sensitivity/recall, precision, and specificity is shown in Table 7.4. Mod_AlexNet performs better than the baseline AlexNet in DE System in every statistic. Mod_AlexNet outperforms AlexNet by 2.24% in accuracy, which is a significant improvement. Both the precision and sensitivity/recall indicators demonstrate a 2.24% improvement, which is consistent with this improvement. These enhancements imply that, in comparison to AlexNet in DE System, Mod_AlexNet in MA System is much better at accurately classifying both positive cases (sensitivity) and minimising FPs (precision). The specificity statistic demonstrates an impressive 11.59% improvement in the benefit of Mod_AlexNet. This suggests that, as compared to AlexNet in DE System, Mod_AlexNet is especially good at correctly identifying negative instances, lowering the rate of FNs. This is a major gain, particularly in this situation, when accurately identifying the negatives is critical.

Additionally, Mod_AlexNet shows a minor improvement of 1.39% in the F1-Score, which is a balanced measure of precision and recall. This suggests that

Mod_AlexNet achieves a more balanced performance in terms of FPs and FNs by achieving a better balance between precision and recall. To summarise, the percentage of improvement values show that Mod_AlexNet in MA System outperforms AlexNet in DE System in all performance criteria consistently. These enhancements indicate that the changes made to Mod_AlexNet are effective and that it performs better in the given tasks. This thorough performance analysis can direct future MA System optimisation efforts and helps to appreciate the subtle strengths of Mod_AlexNet.

Based on Figure 7.3, utilising SGDM optimiser, it is concluded that for a batch size of 32, AlexNet achieved a training accuracy of 98.50% and a training loss value of 0.05, whereas Mod_AlexNet achieved a lower training accuracy of 97.67% and a training loss of 0.08. Moreover, AlexNet obtained 97.94% training accuracy with a training loss of 0.07, whereas Mod_AlexNet achieved 96.74% training accuracy with a training loss of 0.11 for a batch size of 64. Finally, for a batch size of 512, AlexNet obtained 93.33% training accuracy with a training loss value of 0.2, whereas Mod_AlexNet achieved 91% training accuracy with a training loss of 0.26.

Figure 7.4 shows that utilising Adam optimiser, for a batch size of 32, Mod_AlexNet achieved a training accuracy of 89.94% and a training loss value of 0.39, while AlexNet achieved a lower training accuracy of 89.52% and a training loss of 0.4. Moreover, for a batch size of 64, Mod_AlexNet obtained a 94.01% training accuracy with a training loss of 0.18 for a batch size of 64, while AlexNet achieved an 89.50% training accuracy with a training loss of 0.4. Finally, for a batch size of 512, Mod_AlexNet obtained 90.09% training accuracy and a training loss of 0.28, whereas AlexNet achieved 86.70% training accuracy and a loss of 0.48.

Mod_AlexNet outperformed AlexNet for a batch size of 32, with training accuracies of 93.18% and 87.21% and training loss values of 0.25 and 0.72, respectively, based on Figure 7.5 using RMSProp optimiser. Furthermore, at a batch size of 64, Mod_AlexNet achieved a 93.49% training accuracy with a training loss of 0.25, while AlexNet achieved 87.20% training accuracy with a training loss of 0.69. Finally, with 512 batch size, AlexNet had an improved training accuracy of 86.69% and a training loss of 0.5, but Mod_AlexNet achieved a lower training accuracy of 84.17% and a training loss of 0.52.

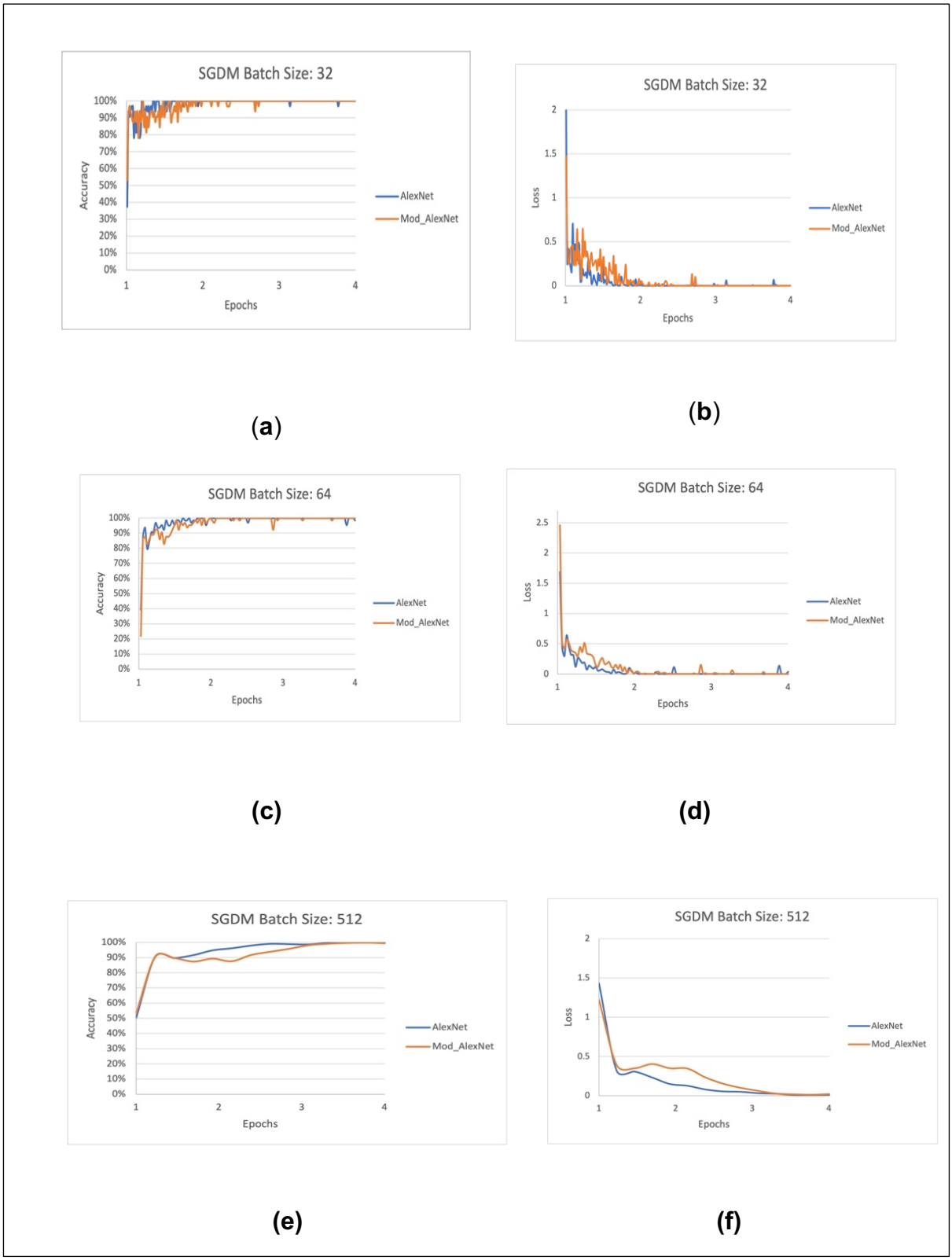

Figure 7.3 *For SGDM optimizer (a) Training Accuracy on Batch size: 32, (b) Training Loss on Batch size: 32, (c) Training Accuracy on Batch size: 64, (d) Training Loss on Batch size: 64, (e) Training Accuracy on Batch size: 512, (f) Training Loss on Batchsize:512*

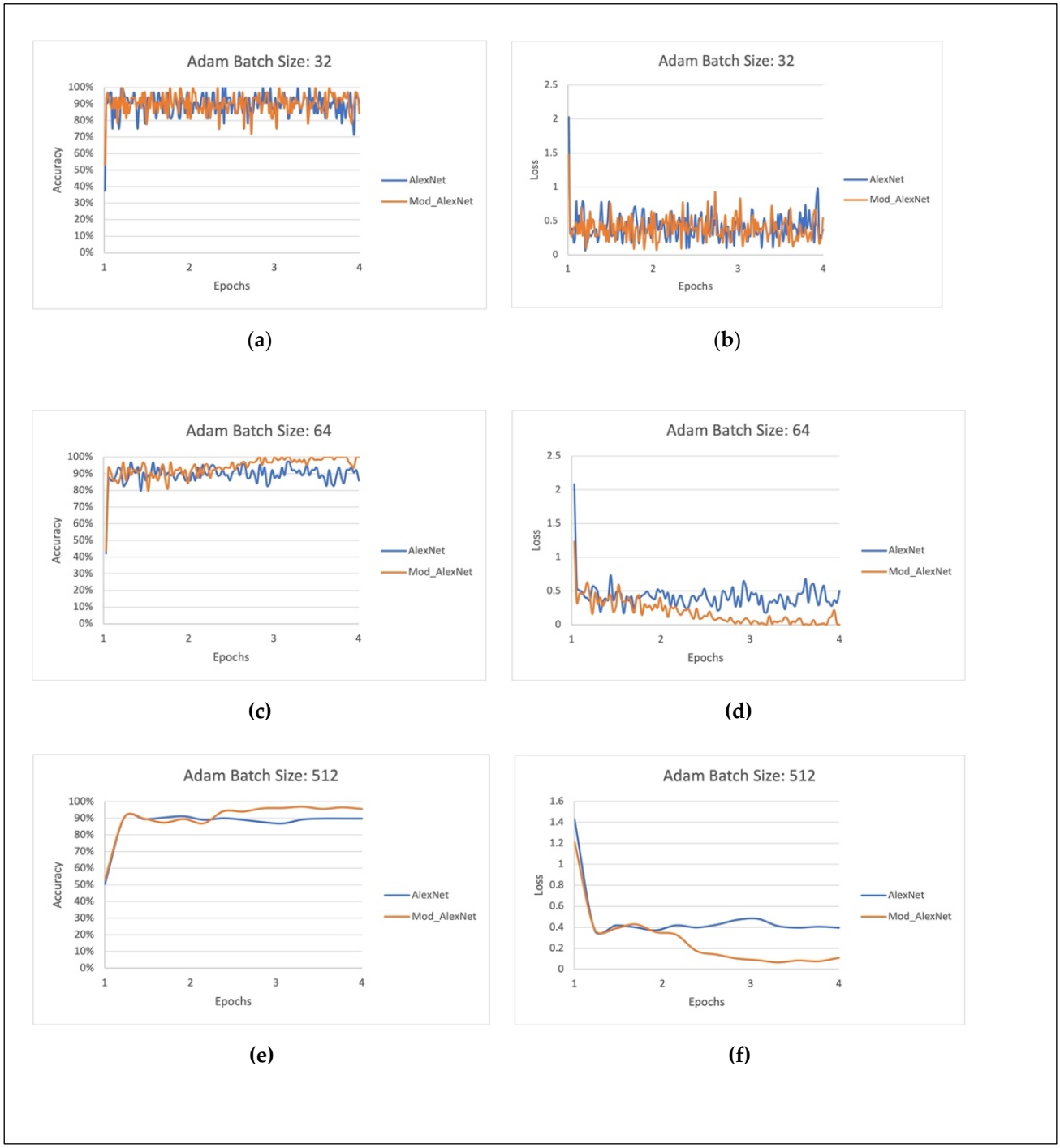

Figure 7.4 *For Adam optimizer (a) Training Accuracy on Batch size: 32, (b) Training Loss on Batch size: 32, (c) Training Accuracy on Batch size: 64, (d) Training Loss on Batch size: 64, (e) Training Accuracy on Batch size: 512, (f) Training Loss on Batchsize:512*

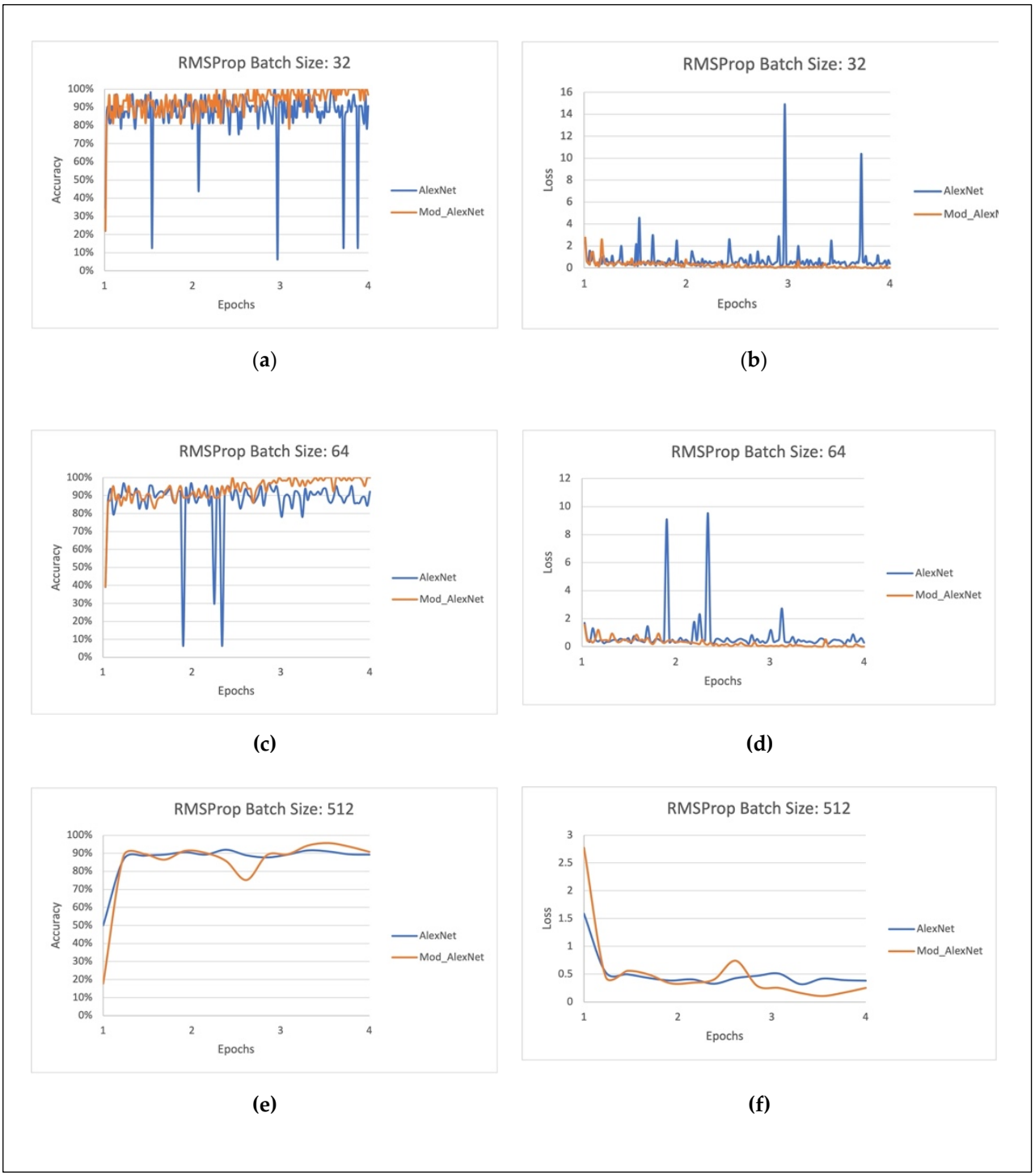

Figure 7.5 *For RMSProp (a) Training Accuracy on Batch size: 32, (b) Training Loss on Batch size: 32, (c) Training Accuracy on Batch size: 64, (d) Training Loss on Batch size: 64, (e) Training Accuracy on Batch size: 512, (f) Training Loss on Batch size: 512.*

While the Mod_AlexNet model demonstrated improvements in accuracy and sensitivity, the consistently low specificity values, particularly when using the Adam and RMSProp optimizers, highlight an important challenge. The tendency of these optimizers to favour sensitivity over specificity may stem from their adaptive learning nature, which adjusts learning rates dynamically and increases the focus on normal cases. The stable but low specificity suggests that the model is more confident in identifying normal cases but struggles with accurately ruling out abnormal cases.

7.4 Summary

In this chapter, MA System was introduced and examined. This builds on the outcomes of the previous chapter, which examined how well cutting-edge deep learning models performed in the classification of DBT images. This chapter presents Mod_AlexNet, a modified deep learning network. The architectural modifications of Mod_AlexNet are essential for addressing the challenges caused by different breast densities and sizes, as well as for accurately differentiating between benign and malignant abnormalities in the identification of breast cancer.

Mod_AlexNet deliberately handles the problem of different breast sizes and densities by adding batch normalisation layers after the first four convolutional layers. These layers guarantee a stable training process by reducing internal covariate shift. Given the variety of structures present in varying breast sizes, internal covariate shift is especially relevant when discussing the diagnosis of breast cancer. Batch normalisation ensures that features are distributed uniformly throughout the network, which promotes reliable learning. This stability is essential for effective training, especially in deep networks such as AlexNet where stability is increasingly difficult to maintain. Early normalisation reduces sensitivity to changes in input data and improves the specificity and generalisation of the model in identifying instances with different breast densities by ensuring that inputs to later layers have a constant distribution. Furthermore, the difficulties in differentiating between benign and malignant abnormalities are addressed by the positioning of max-pooling layers after the third and fourth convolutional stages in Mod_AlexNet. The detection of spatial hierarchies in the incoming data is aided by these layers. Located at these pivotal layers, the model gradually catches and understands complex patterns, leading to a deeper and more selective feature representation. The extraction of hierarchical

features is essential for identifying increasingly complex patterns in the data, which in response improves classification accuracy. Moreover, by lowering the chance of overfitting, a critical factor in handling the complex properties of benign and malignant abnormalities, the down sampling effect of the max-pooling strengthens the model.

In the results and discussion section, the effectiveness of Mod_AlexNet in MA System is thoroughly analysed in comparison to the conventional AlexNet. A variety of optimisers (SGDM, Adam, RMSProp) and batch sizes are evaluated, considering measures like F1-Score, accuracy, sensitivity, precision, and specificity. The results draw attention to important compromises, such as how batch size affects specificity, sensitivity, and precision. Significantly, Mod_AlexNet performs better, particularly when using the SGDM optimiser and a batch size of 32. The chapter uses statistical tools to evaluate the significance of the observed improvements, such as T-tests and Cohen's d. The findings highlight the practical relevance of performance disparities and stress the need to take impact sizes into account in addition to statistical significance. Additionally, a comparison with the most effective outcomes of the DE System is provided, demonstrating the outperformance of Mod_AlexNet. The chapter ends with a thorough analysis of the performance metrics, highlighting the system-wide improvements it continually achieves over AlexNet.

Analysing the performance of AlexNet in both DE System and MA System revealed that using AlexNet for feature extraction and then classifying the extracted features using an SVM classifier achieved more effective outcomes than using AlexNet for feature extraction and classification alone. While the Mod_AlexNet model achieved improved performance across all metrics, specificity remained a challenge. This system laid the groundwork for addressing the challenges caused by varying breast densities and the difficulty in differentiation between benign and malignant tumours but highlighted the need for more advanced methods to further enhance specificity. FFS-EC is developed in response to this observation. Enhancing specificity, and addressing the fundamental challenges in classifying cases of abnormalities, were the main motivations behind the development of FFS-EC. The following chapter describes the development and improvements made to FFS-EC.

Chapter 8 Feature Fusion and Selection with Ensemble Classifier (FFS-EC) System

8.1 Introduction

In the previous chapter, a system comprising Mod_AlexNet model is explored, which was built on the findings of the initial comparative evaluation system. The system introduced in Chapter 6 investigated the classification performance of state-of-the-art deep learning models utilising DBT images. A modified deep learning network named Mod_AlexNet was presented in Chapter 7, with the goal of improving detection accuracy and optimising the identification of abnormal cases. The AlexNet architecture has been modified with more layers, such as max-pooling and batch normalisation, to improve feature extraction, stability, and training effectiveness. While Mod_AlexNet yielded improved outcomes, the primary motivation for establishing this system was to improve specificity and address basic issues involved with the classification of abnormal cases.

To address the previously outlined challenge of correctly classifying abnormal cases, FFS-EC System was developed. The goal of this system was to provide an enhanced framework for the classification of DBT data through the integration of deep learning models with feature fusion, selection and classification ensemble models. To enhance the extraction of prominent features, the suggested system combines the HOG with the HSV colour scheme. The dual application of feature fusion and selection models improves DBT scan breast lesion discrimination. For training the DBT dataset, FFS-EC uses two pre-trained models, ResNet-50 and SqueezeNet, in addition to the previously developed Mod_AlexNet deep learning model. After extracting features from the deep learning models, a series of fusion and selection processes are applied one after the other. The features that were selected are subsequently classified using several classifiers. The final model in this system, an ensemble classifier, combines the predictions of multiple classifiers, attempting to improve classification performance across different measured performance measures.

8.2 Methodology

Within this system, features were extracted once from the pre-trained models ResNet-50 and SqueezeNet, as well as the previously developed Mod_AlexNet. Then, for each scenario, these extracted features were fused with HOG descriptors. HOG creates histograms of gradient orientations for each of the small cells it divides from the image, after which it computes the gradients (directional intensity changes) within each cell. The distribution of edge directions is captured by these histograms, which offer a unique depiction of the local structure and texture in the image. The ability of HOG descriptors to accurately describe object shape and structure while remaining invariant to variations in illumination and contrast is one of the main benefits of extracting them in this system. Concerning the classification of DBT scans, HOG descriptors provide obvious advantages in accurately representing the shape and structure of breast tissue. Their ability to accurately represent the complex spatial details of breast tumours and structures yields a distinct and stable representation that is essential for detecting abnormalities in DBT images.

A variety of models, including Mod_AlexNet, SqueezeNet, and ResNet-50, were implemented after the HSV colour space model was integrated to extract image features. These models were used in conjunction with the use of HOG descriptors. The goal of this multimodal approach was to utilise the advantage of unique capabilities of each model in creating a more complete and complex representation of the data that is found within the images.

The system diagram for this system, the second contribution, is illustrated in Figure 8.1. In Chapter 3, Section 2, an in-depth exposition of the methodologies employed within this system is presented.

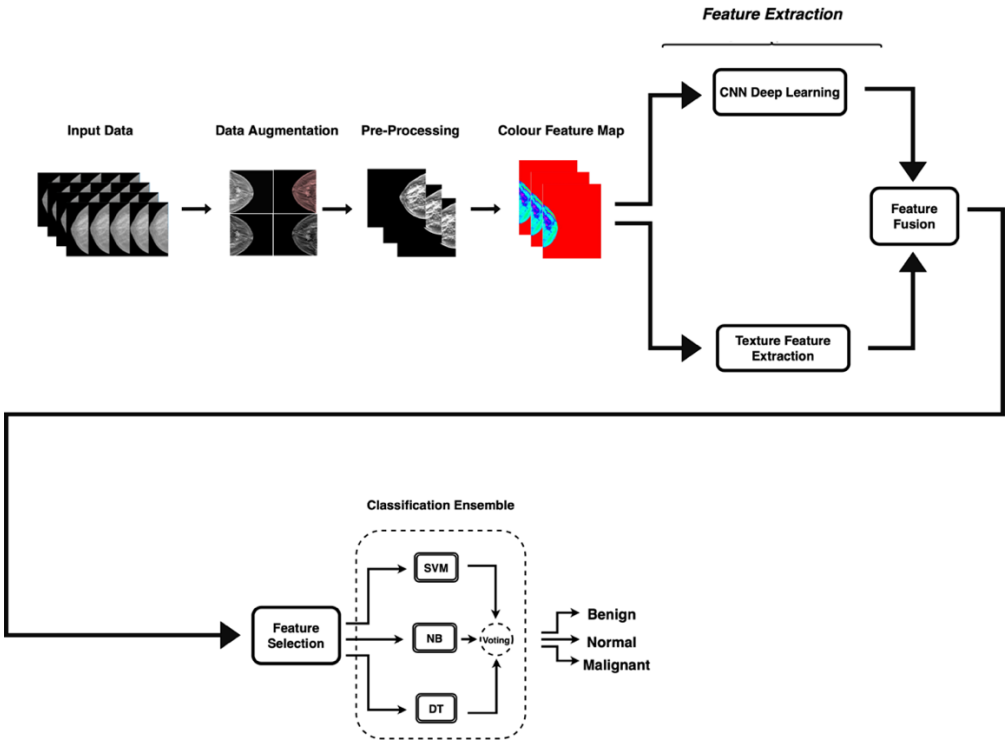

Figure 8.1 *Third Proposed System: An Enhanced System with Feature Fusion and Selection Integrating a Majority Voting Ensemble Classification Model*

Mod_AlexNet, an improved version based on the original AlexNet model, performed well, showing increased effectiveness over the traditional AlexNet. The modifications introduced in Mod_AlexNet, including the addition of max-pooling and batch normalization layers, were carefully designed to enhance training stability and improve feature extraction efficiency. The batch normalization layers address internal covariate shift by ensuring consistent input distributions across layers, thereby facilitating improved gradient flow and accelerating convergence during training. The max-pooling layers contribute to reducing overfitting by down-sampling feature maps, which enhances the model's ability to generalize and effectively capture spatial hierarchies in breast tissue patterns. The increased complexity resulting from these modifications is justified by the observed improvements in classification accuracy and model stability demonstrated through comparative evaluations.
This advantage was most noticeable when it came to the careful classification of DBT images, especially those that belonged to abnormal classes. As a result, features were extracted from the "fc7" layer of Mod_AlexNet, an intentional choice that was required due to this the exceptional ability of the layer to reveal the most

discriminative features present in images. The "fc7" layer is well-known for its skill in extracting abstract and high-level features, which enhances the depth of data for further analysis and interpretation.

The features were extracted from SqueezeNet, an efficient deep learning network that is known for using significantly less parameters than other models to achieve competitive classification accuracy. The architecture of SqueezeNet includes unique 1x1 convolutions and fire-modules that improve the ability of the network to capture complex input and increase prediction accuracy (Iandola et al., 2016). In order to maximise the significance of the extracted features for further classification tasks, special attention was given to the features that came from the "Fire9" layer. This layer is recognised for its distinct structure and functionality, which significantly increases the ability of the extracted features to discriminate, improving the overall performance (Li et al., 2021).

The concept behind using ResNet-50 in this study is the effective application of residual learning, a method that effectively tackles the vanishing gradient issue. This breakthrough makes it possible to build deeper neural networks without compromising performance. Strategic use of skip connections in ResNet-50 enhances its capacity to train enormously deep networks and makes it easier to fully capture complex features and patterns (He et al., 2016). This addition makes an important contribution to the overall improvement in performance. The global average pooling (GAP) layer of the ResNet-50 architecture extracted significant features that were obtained through the extraction process, providing a detailed and complex representation of the images (Zhang et al., 2023).

Features extracted from the deep learning model were fused with HOG using concatenation, to combine the representational abilities of deep neural networks with the advantages of conventional approaches. The HOG features have demonstrated efficacy in image classification by capturing information about local gradients and edge orientations. Integrating these handcrafted features with the high-level and abstract features extracted from the deep learning model results in a fusion that takes advantage of both low-level and details. Rich contextual information and abstract representations are captured by the last layer of the deep learning models, which serves as a feature extractor in this fusion technique. A key component of this

system are the deep learning models implemented, including ResNet, SqueezeNet and Mod_AlexNet, each of which has a unique capacity to recognise complex patterns. Concatenation of these diverse feature sets results in a hybrid feature representation. This combination makes it easier to understand the visual content in the images more comprehensively and improves the robustness of the model, discriminative ability, and generalisation. A comprehensive feature space is generated by integrating HOG features with those derived from deep learning. This approach improves the overall performance metrics through the generation of more accurate predictions.

Following the concatenation of features, feature selection techniques were employed to reduce the number of features assigned as input for classifiers. By maintaining relevant and informative features and removing redundant or less discriminative ones, this model can be used to reduce the influence of dimensionality and improve model efficiency. This process reduces the computational burden of the model and improves its capacity to handle a variety of input features, which in turn increases the overall efficacy of the classifiers. mRMR, the chi-square test, and the f-test were the three different feature reduction techniques applied.

A crucial step in fine-tuning the integrated feature set after feature fusion is the addition of mRMR. The goal of the mRMR goal is to carefully pick a subset of features that balance high relevance to the target variable with low inter-feature redundancy. To determine which features are most relevant to the classification task, the algorithm starts by giving relevance scores to each individual feature depending on its relationship with the target variable. In order to quantify the information overlap, mRMR simultaneously computes redundancy measures between pairs of features. Next, the algorithm ranks features iteratively, choosing at each stage those with the highest relevance and the least amount of repetition. In addition to eliminating redundancy, which might limit model interpretability and generalisation, this method guarantees the inclusion of features essential for precise classification. The algorithm then chooses a collection of features with the lowest redundancy and maximum relevance. This will ensure that the selected features are non-redundant and informative, hence improving the accuracy and efficiency of subsequent classification.

One statistical technique used to evaluate the degree of independence between individual features and the target class is the chi-square test for feature selection. This test examined whether there is a significant correlation between each attribute and the class labels using categorical data. The test determined the difference between the observed and predicted frequencies of feature-class pairings by computing a chi-square statistic. A stronger link is shown by a greater chi-square score. Evaluating the relevance of the fused features to the target variable is part of integrating the chi-square test after the feature fusion stage. This approach assisted in choosing the most useful features for further classification tasks by highlighting features that show strong correlations with class labels. The chi-square test can identify features that greatly increase the discriminative capability of the model and improve its predictability and interpretability in situations where feature fusion has been implemented.

One of the most significant developments in this system has been the addition of an ensemble classifier. This allows us to combine the predictive power of several different algorithms, including DT, NB, and SVM. Using an ensemble methodology is driven by the differences between these classifiers, as each has distinct advantages and disadvantages. The system aimed to overcome the constraints of individual algorithms and produce a more reliable and accurate prediction system by combining their outputs. The SVM, NB, and DT classifiers were carefully configured as part of our ensemble architecture to guarantee the best possible integration. The fusion methodology utilised the benefits of the voting process, an ensemble approach that reduces overfitting risk and improves the performance.

Utilising the combined power of SVM, DT, and NB when trained on the same dataset, the Voting Ensemble classification method is a robust technique. SVM, DT, and NB are the three basis classifiers that each train separately on the training data to identify different patterns within the same overall dataset. The efficacy of the Voting Ensemble is derived from the fusion of predictions produced by every classifier. The Voting Ensemble operates by combining predictions from the base classifiers to arrive at a final prediction. The outputs of SVM, DT, and NB are merged, aligning their different viewpoints, via an ensemble voting process. The class that gets the most votes among the base classifiers is selected as the final

prediction in a hard voting process. By doing this, the danger of overfitting is reduced, and the individual advantages of each classifier are used to improve overall prediction performance. Together, the ability of SVM to create high-dimensional separations of feature space, the capability of DT in capturing complex decision structures, and the probabilistic reasoning of NB result in a more accurate and robust model that skilfully handles the complexities seen in the training dataset across the classes. During the training and testing phases, the ensemble classifier performed better than the individual classifiers on a number of metrics, such as accuracy, precision, and recall. Findings for FFS-EC will be outlined and examined in the subsequent subsection.

8.3 Results and Discussion

This system includes three different scenarios established, each using a different deep learning model for feature extraction. To evaluate and compare the performance of state-of-the-art models, a variety of deep learning models were used. This section explains the use of three deep learning models—Mod_AlexNet, SqueezeNet, and ResNet-50—particular to the feature extraction phase in the corresponding scenarios. Extracted features from each deep learning model were concatenated together with HOG descriptors and then reduced using three different feature selection methods. The final phase involved feeding the selected features into the classifiers and aggregating them using the voting ensemble model.

To mitigate the risk of overfitting associated with ensemble models, cross-validation was used to evaluate the model's ability to generalize to unseen data. Additionally, feature reduction techniques, including mRMR, chi-square, and f-test, were applied to eliminate irrelevant or redundant features, improving model efficiency and reducing the likelihood of overfitting. This approach helped maintain consistent performance across both training and validation sets, enhancing the model’s overall robustness.

8.3.1 SqueezeNet

In this scenario, "Fire9" layer features were extracted and concatenated with HOG descriptors. Feature selection methods (namely: mRMR, chi-square test, and f-test) were then utilised to select robust features and reduce the dimensionality of the feature vector. The selected features were then fed into three different classifiers

(SVM, NB, and DT), and the final prediction was obtained through integrating the outputs of each classifier with a voting ensemble classification model. An analysis of the results and the calculation of different performance indicators were computed. The results are displayed in Table 8.1 for the first scenario.

A comprehensive overview of the performance metrics of SqueezeNet classifier over a range of integrated instances, including different classifiers and feature selection techniques, is given in Table 8.1. A thorough examination of the effects of many factors on the performance is made possible by the ability to represent each row as a distinct configuration. In the initial instance employing only SqueezeNet features, the SVM classifier attained an accuracy of 82.68%. Despite a high sensitivity of 82.68%, precision was 86.72%, indicating a significant percentage of FPs. The NB classifier had lower accuracy (46.73%), but greater precision (85.62%). With an accuracy of 83.23%, DT fared well, exhibiting a balance between sensitivity, precision, and F1-Score.

Table 8.1 The performance of SqueezeNet-based system with different integrated subphases.

Different integrated contexts	Classifier	Performance Measures				
		Accuracy	Sensitivity	Precision	Specificity	F1-Score
SqueezeNet only	SVM	82.68%	82.68%	86.72%	28.60%	84.55%
	NB	46.73%	46.73%	85.62%	54.69%	59.83%
	DT	83.23%	83.23%	86.80%	29.27%	84.92%
SqueezeNet and Feature Fusion	SVM	82.34%	82.34%	86.85%	30.86%	84.42%
	NB	47.24%	47.24%	85.92%	56.21%	60.02%
	DT	83.33%	83.33%	86.92%	30.17%	85.02%
SqueezeNet, Feature Fusion, and mRMR	SVM	92.27%	92.27%	91.07%	32.32%	91.12%
	NB	48.73%	48.73%	86.34%	57.78%	61.25%
	DT	84.11%	84.11%	87.35%	32.01%	85.64%
SqueezeNet, Feature Fusion, and chi-square test	SVM	90.12%	90.12%	88.23%	26.14%	89.09%
	NB	48.56%	48.56%	86.23%	57.17%	61.19%
	DT	83.99%	83.99%	87.22%	31.40%	85.51%
SqueezeNet, Feature Fusion, and f-test	SVM	90.38%	90.38%	88.28%	26.15%	89.22%
	NB	48.28%	48.28%	86.11%	56.56%	60.95%
	DT	83.78%	83.78%	87.25%	32.00%	85.42%
Ensemble for mRMR		92.78%	92.78%	91.58%	35.36%	91.70%
Ensemble for chi-square test		90.62%	90.62%	89.26%	32.87%	89.87%
Ensemble for f-test		90.88%	90.88%	89.40%	32.88%	90.05%

The SVM classifier demonstrated an impressive accuracy of 82.34% when SqueezeNet features and HOG descriptors were integrated. The sensitivity of this accuracy reflected the balanced capacity of SVM to accurately recognise positive events. The accuracy in detecting positive instances among its predictions was demonstrated by its precision of 86.85%. Specificity demonstrated a small capacity to correctly identify negative instances, while being low at 30.86%. With an F1-Score of 84.42%, SVM demonstrated its overall ability to maintain a balanced ratio of sensitivity to precision. The accuracy of the NB classifier was 47.24%, which was lower than that of SVM. At 47.24%, sensitivity and accuracy were in line, suggesting a moderate capacity for accurately identifying positive events. On the other hand, high precision (85.92%) of NB, highlighted its ability to correctly detect positive situations. NB stood out for having a high specificity of 56.21%, which means that it was quite good at recognising negative cases. With an F1-Score of 60.02%, the performance was balanced in terms of sensitivity and precision. With an accuracy of 83.33%, the DT classifier outperformed both SVM and NB. Accuracy and sensitivity matched, suggesting a strong capacity to recognise good examples. With a precision of 86.92%, the classifier demonstrated its ability to correctly identify positive cases. With a specificity of 30.17%, the capacity to accurately detect negative cases was considered to be low. With an F1-Score of 85.02%, the performance was well-balanced between sensitivity and precision. After the comparative study, the DT classifier was found to be the most effective surpassing both SVM and NB in terms of accuracy, precision and f1-score. Both DT and SVM outperformed NB in terms of precision, suggesting better detection of positive cases. When compared to SVM and DT, the significant superiority of NB in specificity indicates a greater capacity to accurately identify negative events. With the greatest F1-Score, the DT classifier performed exceptionally well, demonstrating a balance between sensitivity and precision.

The investigation of integrating SqueezeNet features with HOG descriptors and then selecting mRMR features involves an in-depth examination of three classifiers: DT, NB, and SVM. When combined with mRMR feature selection using SqueezeNet and HOG descriptors, the SVM classifier achieved an outstanding accuracy of 92.27%. The strong ability of SVM to accurately detect positive instances was highlighted by the high accuracy that was reflected in the sensitivity. With a precision of 91.07%,

the classifier demonstrated its ability in finding positive examples within its predictions. Despite a modest specificity of 32.32%, suggesting a limited ability to properly detect negative cases, the F1-Score reached an outstanding 91.12%, showcasing the general skill of SVM in maintaining a harmonious balance between precision and sensitivity. The NB classifier, on the other hand, showed a lower accuracy of 48.73%, indicating its limits when compared to SVM. With a notable specificity of 57.78%, NB was exceptionally good at correctly detecting negative cases. The DT classifier performed better than NB but not quite as well as SVM, with a high accuracy of 84.11%. The accuracy in identifying positive cases was demonstrated by its precision of 87.35%. With a specificity of 32.01%, the capacity to accurately detect negative cases was deemed to be low. To summarise, SVM was found to be the most effective classifier, outperforming NB and DT in terms of accuracy, precision, and F1-Score. The DT classifier proved to be a competitive alternative with an impressive overall performance, especially in precision and F1-Score. Although accuracy of NB was restricted, its specificity was excellent, making it an appropriate choice in some situations.

In the fusion of SqueezeNet features and HOG descriptors, coupled with chi-square feature selection, the performance of three classifiers—SVM, NB, and DT—was examined. SVM proved to be effective in accurately identifying situations, as evidenced by its exceptional accuracy of 90.12%. In addition, the classifier demonstrated good precision (88.23%) and sensitivity (90.12%), demonstrating its capacity to correctly identify positive cases while reducing FPs. In contrast, NB yielded lower accuracy at 48.56%, but excelled in specificity (57.17%) and attained an adequate F1-Score of 61.19%. The DT classifier performed well, achieving 83.99% accuracy, 83.99% sensitivity, and 87.22% precision. At 31.40%, its specificity was, nevertheless, comparatively poor. With a balance between sensitivity and precision, SVM yielded the highest F1-Score (89.09%), closely followed by DT (85.51%) and NB (61.19%). The SVM is the most efficient classifier in terms of accuracy, precision, and f1-score; DT produces results that are competitive, whereas NB performs better in terms of specificity. The performance of three classifiers—SVM, NB, and DT—was assessed in the integration of SqueezeNet features and HOG descriptors using f-test feature selection. With an accuracy of 90.38%, SVM proved that it could accurately categorise occurrences. In addition, the classifier

demonstrated good precision (88.28%) and sensitivity (90.38%), indicating its ability to correctly identify positive cases while reducing FPs. In contrast, the accuracy of NB was lower at 48.28%, but its specificity was outstanding at 56.56%, and its F1-Score weighed in at 60.95%. The DT classifier yielded an 83.78% accuracy rate, along with noteworthy precision (87.25%) and sensitivity (83.78%). At 32.00%, its specificity was, nevertheless, significantly lower. SVM has the highest F1-Score (89.22%), closely followed by DT (85.42%) and NB (60.95%).

Across a range of integrated contexts, the ensemble methods that employ distinct feature selection methodologies demonstrated noteworthy performance in identifying occurrences. With an accuracy of 92.78%, the mRMR-based ensemble showed the highest level of proficiency in accurate classification. Meanwhile, the chi-square and f-test ensembles earned accuracies of 90.62% and 90.88%, respectively, displaying their competitive capabilities. The mRMR ensemble continuously maintained high values at 92.78% and 91.58% when sensitivity and precision were examined, demonstrating robust detection of positive cases with minimised FPs. With sensitivity and precision values of 90.62% and 89.26% and 90.88% and 89.40%, respectively, the chi-square and f-test ensembles demonstrated efficient positive instance identification while maintaining precision.

As compared to the baseline of classifying features extracted from SqueezeNet and based on results from Table 8.1, the results offer a comprehensive analysis of the percentage improvement in model performance. The given results provide an in-depth analysis of the percentage gain in model performance over the baseline of classifying features that were taken only from SqueezeNet. SVM shows a tiny drop of -0.41% in accuracy, sensitivity, and F1-Score in the integration of SqueezeNet features with HOG descriptors, with an increase in specificity (7.90%), suggesting generally constant performance. Notably, NB shows minor gain in F1-Score (0.32%) and accuracy (1.09%), while DT displays slight enhancements. In the SqueezeNet, Feature Fusion, and mRMR setup, SVM displays a large improvement of 11.60% across all measures, emphasising the importance of mRMR in increasing the overall performance of SVM. NB demonstrates significant gains in F1-Score (2.37%), sensitivity (4.28%), and accuracy (4.28%), demonstrating the beneficial impact of mRMR in capturing important features. DT exhibits moderate advances in accuracy,

sensitivity, and F1-Score, highlighting the beneficial effect of mRMR on its performance. When SVM switches to SqueezeNet, Feature Fusion, and chi-square test, it records significant gains in accuracy (9.00%) and F1-Score (5.37%), demonstrating how well chi-square feature selection works to increase performance overall. But a drop in specificity (-8.59%) points to a compromise. NB shows increase in F1-Score (2.27%) and accuracy (3.91%), indicating a favourable effect on sensitivity balance and precision. In this integrated setting, DT demonstrates stability with slight improvements. SVM shows notable gains in accuracy (9.31%) and F1-Score (5.52%) in SqueezeNet, Feature Fusion, and f-test, suggesting that f-test feature selection has a beneficial effect on accurate classification, even while specificity (-8.56%) decreases. The mild improvements that NB experiences indicate that the f-test has a good impact on the performance of NB, as evidenced by the balanced increases in F1-Score, accuracy, and sensitivity. In this integrated setting, DT exhibits minor improvements, indicating consistent performance. When considering the ensemble methods, Ensemble for mRMR shows substantial improvements in all measures, especially in specificity (20.79%) and F1-Score (7.99%), demonstrating how well mRMR feature selection works when combined with ensemble model. The ensemble for the chi-square test shows significant gains in F1-Score (5.83%), sensitivity (8.88%), and accuracy (8.88%). The f-test ensemble exhibits noteworthy increases in all measures, particularly in specificity (12.33%) and F1-Score (6.04%), indicating an important part of f-test feature selection in the ability of the ensemble to recognise positive occurrences.

Two measures were used to evaluate the relevance and importance of the enhancements: the t-test and Cohen's measure. In the comparison, the performance of the classifiers solely using SqueezeNet features—i.e., without feature fusion, selection, or ensemble classification models—is compared to their significance. The Cohen d's measure and t-test results for each situation are shown in Table 8.2.

Using the Cohen's d measure and significant values, the reported findings thoroughly assess the improvement in model performance relative to the baseline of classifying features that were extracted from SqueezeNet. SVM, NB, and DT produced small statistically significant Cohen's d values (0.012965, 0.037966, 0.010737) in the feature fusion context with HOG descriptors, indicating relatively minor practical

significance. After switching to SqueezeNet, Feature Fusion, and mRMR, SVM showed a medium Cohen's d value of 0.26277, indicating a statistically significant improvement with a moderate practical significance. Though statistically significant, the observed Cohen's d values for NB and DT were tiny (0.1173, 0.047448), suggesting comparably little practical impact. SVM, NB, and DT showed tiny Cohen's d values (varying from 0.038116 to 0.14363) in the SqueezeNet, Feature Fusion, and chi-square test and SqueezeNet, Feature Fusion, and f-test, showing minor but statistically significant improvements with little practical impact.

Table 8.2 Comparative Statistical Analysis using Cohen's d and T-test Significance for Evaluating Significance of Improvement against the Performance of Classifiers when Classifying Features Extracted from SqueezeNet.

Improvement compared to when classifying features extracted from SqueezeNet only	Classifier	Cohen's d:		T-test (p)	
		Measure	Significance	Measure	Significance
SqueezeNet and Feature Fusion	SVM	0.013	Small	0.5569	Not Significant
	NB	0.038	Small	0.0626	Not Significant
	DT	0.011	Small	0.1723	Not Significant
SqueezeNet, Feature Fusion, and mRMR	SVM	0.263	Medium	0.0056	Significant
	NB	0.117	Small	0.0091	Significant
	DT	0.047	Small	0.0452	Significant
SqueezeNet, Feature Fusion, and chi-square test	SVM	0.139	Small	0.1221	Not Significant
	NB	0.103	Small	0.0063	Significant
	DT	0.038	Small	0.0384	Significant
SqueezeNet, Feature Fusion, and f-test	SVM	0.144	Small	0.1192	Not Significant
	NB	0.083	Small	0.0052	Significant
	DT	0.039	Small	0.0968	Not Significant
Ensemble for mRMR		0.293	Medium	0.0015	Significant
Ensemble for chi-square test		0.205	Medium	0.0066	Significant
Ensemble for f-test		0.211	Medium	0.0066	Significant

All three of the ensemble models that included the f-test, chi-square test, and mRMR showed medium Cohen's d values (0.20488 to 0.29276). This points to a more significant practical significance and significant increases in the performance of the model. To illustrate the findings, the ensemble model with mRMR is the most significant, with a medium Cohen's d value of 0.29276. This indicates that the

performance of the model has improved in a statistically and practically significant way. The inclusion of mRMR in the ensemble significantly improves F1-Score, sensitivity, specificity, accuracy, and precision, demonstrating how well it works to maximise TPs while reducing FPs in the ability of the classifier to identify positive cases. To sum up, the ensemble model that integrates mRMR proves to be the most efficient and holds significant practical value. This emphasises the value of carefully choosing features and integrating them, highlighting the applicability of mRMR in enhancing overall model performance.

The results that are shown here focus on evaluating the improvement in comparison to the classification of features that are only derived from SqueezeNet. T-test (p) values are used in this assessment to determine the importance of each classifier and feature integration technique. In the instance of combining SqueezeNet features with HOG descriptors, SVM, NB, and DT classifiers exhibit T-test (p) values of 0.5569, 0.0626, and 0.1723, respectively, indicating that the improvements are not statistically significant. The three classifiers—SVM, NB, and DT—show T-test (p) values of 0.0056, 0.0091, and 0.0452, respectively, after switching to the SqueezeNet, Feature Fusion, and mRMR configuration. The statistical significance of the low p-values indicates that the observed improvements are probably the result of the feature fusion and selection procedures that were used, rather than being the result of chance. NB and DT classifiers consistently demonstrate statistically significant improvements with low T-test (p) values for the SqueezeNet, Feature Fusion, and chi-square test and SqueezeNet, Feature Fusion, and f-test configurations, demonstrating the efficacy of these feature selection techniques. Notably, T-test (p) values of 0.0015, 0.0066, and 0.0066, respectively, suggest significant improvements in the ensemble methods: Ensemble for mRMR, Ensemble for chi-square test, and Ensemble for f-test. The ensemble models add to a combined improvement in model performance by integrating the efficacy of individual classifiers and feature selection strategies. T-test (p) results, in summary, highlight the statistical significance of the observed gains using specific feature integration techniques, especially when considering mRMR and ensemble models. These results offer important new information on which strategies work best for improving the performance.

8.3.2 ResNet-50

In this scenario, HOG descriptors were concatenated with features recovered from the ResNet-50 "GAP" layer. Afterwards, a number of feature selection methods were utilised to select robust features and reduce the dimensionality of the feature vector, such as mRMR, chi-square test, and f-test. The final prediction was obtained by combining the outputs of each classifier using a voting ensemble classification model. The selected features were then fed into three different classifiers (SVM, NB, and DT). For the second scenario, the outcomes of the evaluation process and the computation of various performance indicators are shown in Tables 8.3 and 8.4.

Table 8.3 The performance of ResNetNet-50-based developed system with different integrated subphases.

Different integrated contexts	Classifier	Performance Measures				
		Accuracy	Sensitivity	Precision	Specificity	F1-Score
ResNet-50 only	SVM	85.35%	85.35%	86.38%	22.22%	85.84%
	NB	45.98%	45.98%	86.49%	60.91%	58.84%
	DT	84.19%	84.19%	86.51%	24.89%	85.30%
ResNet-50 and Feature Fusion	SVM	89.49%	89.49%	87.74%	22.91%	88.46%
	NB	48.81%	48.81%	87.16%	62.85%	61.30%
	DT	84.20%	84.20%	87.12%	31.03%	85.59%
ResNet-50, Feature Fusion, and mRMR	SVM	90.65%	90.65%	88.82%	27.06%	89.61%
	NB	50.37%	50.37%	86.61%	57.98%	62.60%
	DT	84.09%	84.09%	86.81%	28.47%	85.35%
ResNet-50, Feature Fusion, and chi-square test	SVM	88.91%	88.91%	87.88%	26.40%	88.37%
	NB	51.11%	51.11%	87.35%	62.51%	63.10%
	DT	83.66%	83.66%	86.81%	28.75%	85.12%
ResNet-50, Feature Fusion, and f-test	SVM	89.39%	89.39%	87.97%	25.51%	88.63%
	NB	51.31%	51.31%	87.29%	61.92%	63.28%
	DT	82.04%	82.04%	86.31%	26.59%	84.06%
Ensemble for mRMR		87.14%	87.14%	89.07%	32.64%	87.77%
Ensemble for chi-square test		86.61%	86.61%	88.16%	29.61%	87.16%
Ensemble for f-test		85.54%	85.54%	87.49%	28.97%	86.40%

The output of the system, which was obtained by extracting ResNet-50 features and feeding them into several classifiers, shows unique performance traits in various integrated scenarios, as shown in Table 8.3. The SVM classifier had the highest

accuracy at 85.35%, proving that the extracted ResNet-50 features were effectively utilised by SVM for classification. Nevertheless, a significant disadvantage is apparent in the specificity, which is only 22.22%. The SVM classifier performed well in detecting positive cases but had difficulty correctly identifying negative examples, as suggested by the high sensitivity and precision.

The accuracy of the NB classifier, on the other hand, was noticeably lower at 45.98%. The inability to handle the complexity of the ResNet-50 features is the reason for the lower accuracy. Although the precision rate of 86.49% of NB is rather good, its sensitivity and specificity are degraded, suggesting that it may not be able to categorise both positive and negative cases with sufficient accuracy. With an accuracy of 84.19%, the DT classifier demonstrated competitive performance by balancing sensitivity and precision. But just like SVM, DT has trouble reaching a high enough specificity, illustrating how difficult it is to recognise negative cases in the context of classifying features extracted from ResNet-50. The common challenge in utilising deep features for classification is shown by the observed limitations in specificity across classifiers in the ResNet-50 alone scenario. Although the extraction of features from ResNet-50 can be advantageous in some classification scenarios, it may present inherent complications that classifiers must deal with. Even though the SVM classifier has the highest accuracy, it has trouble distinguishing negative cases, which may mean that the way ResNet-50 features are handled needs to be improved.

Upon system evaluation, significant performance characteristics are revealed. ResNet-50 features were extracted and fused with HOG descriptors using concatenation, followed by classification utilising several classifiers. With an accuracy of 89.49%, the SVM classifier emerges as the best performer. Given its high accuracy, it can be inferred that the SVM model successfully combined the characteristics of ResNet-50 and HOG to achieve correct classification. Significantly, SVM showed high sensitivity and precision, demonstrating its ability to accurately detect positive instances. On the other hand, an important disadvantage is the poor specificity of 22.91%, indicating difficulties in precisely detecting negative cases. In comparison, the NB classifier displayed a lower accuracy of 48.81%, demonstrating difficulty in efficiently using the concatenated features for classification. Although the

precision rate was high (87.16%) of NB, its sensitivity and specificity were weakened, indicating that it may not be able to distinguish positive and negative cases with equal accuracy. With an accuracy of 84.20%, the DT classifier demonstrated competitive performance. DT showed balanced precision and sensitivity, much like SVM. At 31.03%, the specificity continues to have challenges even if it is greater than SVM. The observed differences in classifier performance illustrate how the feature fusion approach affects classification results. SVM was the most effective in employing the fused features to achieve high accuracy, while NB had some limitations, especially with regard to sensitivity and overall accuracy. DT performed competitively, but issues with specificity still exist, highlighting how difficult it is to identify negative instances in the fused feature set. To sum up, concatenating ResNet-50 features with HOG descriptors presents an interesting possibility for improving classification results. Classifier performance variances indicate that classifier selection for fused feature sets should be carefully considered. The high accuracy of SVM shows the effectiveness of this method, but the sensitivity, specificity, and precision trade-offs that have been observed highlight the significance of a balanced assessment of performance metrics when determining the overall effectiveness of the system.

The system analysis, in which ResNet-50 features were extracted and fused with HOG descriptors via concatenation, followed by mRMR feature selection and classification by various classifiers, provides an in-depth understanding of the impact of these processes on different performance measures. Starting with the SVM classifier, it attained a commendable 90.65% accuracy. This high accuracy indicates that the ability of SVM to create accurate predictions was largely influenced by the combination of ResNet-50 and HOG features, followed by mRMR feature selection. At 90.65% and 88.82%, respectively, the sensitivity and precision values show a balanced performance in accurately identifying positive situations. With a specificity of only 27.06%, there are still significant challenges in correctly recognising negative situations. The F1-Score, which analyses the balance between precision and recall, is exceptionally high at 89.61%, emphasising the effectiveness of the classifier in reaching an equilibrium between FPs and FNs. The accuracy is significantly lower for the NB classifier, at 50.37%. This shows that NB has difficulties correctly identifying cases even with feature fusion and mRMR feature selection. With a

precision of 86.61%, NB is likely to be correct when it offers a positive forecast. But the sensitivity and specificity numbers, which are 50.37% and 57.98%, respectively, show how challenging it was for NB to accurately identify both positive and negative situations. The trade-off between recall and precision for NB is highlighted by the F1-Score of 62.60%. With regard to the DT classifier, 84.09% accuracy is attained. Though the specificity of 28.47% still presents a challenge, this shows competitive performance. The balanced precision and sensitivity values of 86.81% and 84.09%, respectively, demonstrate the ability of DT to produce precise positive predictions. Considering the trade-off between precision and recall, the overall performance is boosted by its F1-Score of 85.35%. In conclusion, the SVM classifier performs well overall, especially in terms of accuracy, sensitivity, and precision, but it has challenges with specificity. While the DT classifier performs competitively, the NB classifier struggles with sensitivity and accuracy. The overall accuracy and specificity seem to be positively impacted by the mRMR feature selection method, highlighting the significance of feature selection in fine-tuning the classification results.

This analysis of the system, in which ResNet-50 features were extracted, concatenated with HOG descriptors to fuse them, and then selected chi-square features before being classified by different classifiers, offers an in-depth understanding of the impact of chi-square feature selection model on different performance metrics. The SVM classifier achieved an accuracy of 88.91%, demonstrating strong overall performance. It demonstrated great sensitivity and precision, demonstrating competence in accurately detecting positive cases while preserving prediction accuracy. On the other hand, the issue in specificity, at 26.40%, indicates possible challenges in correctly identifying negative situations. The F1-Score of 88.37% emphasises the ability of the classifier in balancing FPs and FNs. With an accuracy of 51.11%, the NB classifier demonstrated moderate success, especially in precision. Even with a high precision of 87.35%, issues with sensitivity and specificity suggest that a balanced classification of both positive and negative instances may be challenging to achieve. For NB in this context, the balance between recall and precision is highlighted by the F1-Score of 63.10%. 83.66% accuracy was a competitive performance for the DT classifier. It demonstrated its accuracy in positive predictions while retaining precision by achieving a balance between sensitivity and precision. Nonetheless, difficulties with

specificity, at 28.75%, point to difficulties in correctly categorising negative cases. The efficacy of the classifier in striking a balance between recall and precision is demonstrated by its F1-Score of 85.12%. In summary, the SVM classifier excelled with good overall performance, notably in accuracy and sensitivity. The NB classifier demonstrated a moderate level of effectiveness, prioritising precision despite encountering difficulties with sensitivity and specificity. The DT classifier displayed competitive performance but found difficulty in reaching high specificity.

With an accuracy of 89.39%, the SVM classifier exhibited strong performance when the f-test feature selection model was applied. Its ability in precisely and accurately recognising positive instances is demonstrated by the high precision and sensitivity values (89.39% and 87.97%, respectively). On the other hand, an important limitation in specificity—which stands at 25.51%—indicates difficulties in precisely recognising negative cases. The overall efficacy of SVM in striking a balance between FPs and FNs is highlighted by the F1-Score, which stands at 88.63%. In comparison, the NB classifier displayed a lower accuracy of 51.31%, demonstrating difficulty in utilising the f-test-selected features for effective classification. Although the precision rate of NB was high (87.29%), its sensitivity and specificity were impaired (51.31% and 61.92%, respectively), indicating that NB had trouble correctly identifying both positive and negative cases. With an accuracy of 82.04%, the DT classifier demonstrated competitive performance. Based on the balanced sensitivity and precision values of 82.04% and 86.31%, respectively, it can make positive predictions with a high degree of accuracy and precision. On the other hand, difficulties with specificity (26.59%) suggest that it can be difficult to characterise negative situations with accuracy. The overall efficacy of DT classifier in achieving a balance between precision and recall is highlighted by its F1-Score of 84.06%. In conclusion, despite challenges with specificity, the SVM and DT classifiers stand out for their strong overall performance, especially in terms of accuracy, sensitivity, and precision. There are limitations with the sensitivity and accuracy of the NB classifier.

Based on a variety of performance measures, the ensemble classifiers—all of which used unique feature selection methods (mRMR, chi-square test, and f-test)—were carefully assessed. With the mRMR feature selection model, the ensemble classifier showed an overall accuracy of 87.14%. It was noteworthy that it demonstrated great

sensitivity (87.14%) and precision (89.07%), demonstrating its capacity to accurately detect positive cases while preserving accuracy. But the specificity is much lower at 32.64%, indicating difficulties in precisely detecting negative cases. With an F1-Score of 87.77%, recall and precision are harmoniously balanced. The mRMR performance of the ensemble demonstrates how good it is at classifying negative cases overall, but it also emphasises the trade-off between sensitivity and specificity. An accuracy of 86.61% was attained by the ensemble classifier that used the chi-square test for feature selection. It showed balanced precision (88.16%) and sensitivity (86.61%), demonstrating its capacity to make precise positive predictions. The specificity, at 29.61%, is rather lower, though, suggesting that it would be challenging to correctly identify negative examples. The performance of the chi-square test ensemble indicates that it is effective in making accurate positive predictions, but there are still limitations with correctly classifying negative instances, as seen by the lower specificity. Achieving an accuracy of 85.54% was the ensemble classifier that used the f-test for feature selection. Accuracy in positive predictions was maintained while retaining competitive sensitivity (85.54%) and precision (87.49%). The 28.97% specificity indicates challenges with precisely identifying negative cases. The performance of the f-test ensemble highlights how well it can achieve overall accuracy while taking accurate recall and precision into account, however specificity issues still need to be addressed. To summarise, the feature selection models based on mRMR, chi-square test, and f-test ensemble classifiers have complex performance characteristics. The chi-square test prioritises correct positive predictions but has difficulties with specificity, whereas mRMR shows a trade-off between sensitivity and specificity. The f-test ensemble achieves competitive overall accuracy with a balanced consideration of precision and recall, although, like the others, struggles with specificity.

The evaluation of classification performance improvement rates relative to ResNet-50-only classification provides a comprehensive insight into the impact of feature fusion, selection, and ensemble strategies across different classifiers. Based on results from Table 8.3, Slight gains were observed when ResNet-50 features were combined with HOG descriptors for SVM, NB, and DT classifiers. Small accuracy improvements of 4.85% and 6.15% for SVM and NB, respectively, indicate a marginal improvement in accurately detecting cases. The low specificity values, on

the other hand, point to possible inconsistencies between the fused features and the decision limits of the classifiers and suggest difficulties in accurately categorising negative cases. Even though the specificity of the DT significantly improved, its precision and F1-Score suffered, suggesting that there may have been misclassifications and that the overall efficacy of the feature fusion was restricted. Fusing features and applying mRMR feature selection demonstrated a moderate enhancement in SVM accuracy (6.21%), precision, and F1-Score. However, NB encountered challenges, reflected in a negative specificity, suggesting potential misclassifications of negative instances. DT exhibited minor fluctuations, indicating that mRMR had limited impact on its overall performance. The feature selection process improved the ability of SVM to discriminate positive instances, highlighting the compatibility of mRMR-selected features with the decision boundaries of the SVM. Fusing and using chi-square feature selection yielded a variety of results. The accuracy of the SVM improved little (4.17%), but the accuracy of the NB increased significantly (11.15%). The negative specificity for NB, however, raises questions about the difficulty of accurately categorising negative cases. The accuracy of the DT slightly decreased, suggesting that the choice of chi-square features could not have been well-aligned with its decision-making process. These findings highlight how the feature selection process affects the performance of the classifier. For both SVM (4.73%) and NB (11.59%), feature fusion and f-test feature selection led to minor improvements. But DT reported a slight decrease in F1-Score, accuracy, precision, and specificity. The negative specificity values for both NB and DT highlight the possibility of difficulties in accurately classifying negative instances and highlight how the choice of feature selection method and its suitability for the classifiers can have a significant impact on the findings. The inclusion of ensemble models with the chosen attributes showed inconsistent results. Although certain combinations, such as mRMR combined with a voting ensemble, demonstrated outstanding gains in specific measures. When comparing ensemble models that use different feature selection techniques, mRMR outperformed them. These results show that the effectiveness of the ensemble may depend on how well the classifiers and the features selected cooperate.

The t-test and Cohen's measure were the two metrics used to evaluate the significance and impact of the improvements. The assessment compares the

importance of the performance of the classifiers with that of using only ResNet-50 features—that is, without feature fusion, selection, or ensemble classification models. The results of the t-test and Cohen's d measure are shown in Table 8.4 for each case.

In Table 8.4, small effect sizes were found for SVM (0.089), NB (0.133), and DT (0.055) when Cohen's d and t-test were applied. The combination of ResNet-50 features with HOG descriptors resulted in statistically significant improvements in the performance of these classifiers, as evidenced by the t-test, which showed substantial gains for SVM (p=0.0213) and NB (p=0.0060). However, despite a small effect size, DT did not show significant improvement (p=0.3001). With the statistical significance supporting the practical significance for SVM and NB, this difference highlights the subtle effects of feature fusion on various classifiers.

Table 8.4 *Comparative Statistical Analysis using Cohen's d and T-test Significance for Evaluating Significance of Improvement against the Performance of Classifiers when Classifying Features Extracted from ResNet-50.*

Improvement compared to when classifying features extracted from ResNet-50 only	Classifier	Cohen's d:		T-test (p)	
		Measure	Significance	Measure	Significance
ResNet-50 and Feature Fusion	SVM	0.089	Small	0.0213	Significant
	NB	0.133	Small	0.0060	Significant
	DT	0.055	Small	0.3001	Not Significant
ResNet-50, Feature Fusion, and mRMR	SVM	0.153	Small	0.0014	Significant
	NB	0.123	Small	0.2521	Not Significant
	DT	0.029	Small	0.3540	Not Significant
ResNet-50, Feature Fusion, and chi-square test	SVM	0.109	Small	0.0029	Significant
	NB	0.216	Medium	0.0200	Significant
	DT	0.022	Small	0.5219	Not Significant
ResNet-50, Feature Fusion, and f-test	SVM	0.111	Small	0.0023	Significant
	NB	0.216	Medium	0.0299	Significant
	DT	-0.031	Small	0.3262	Not Significant
Ensemble for mRMR		0.145	Small	0.0207	Significant
Ensemble for chi-square test		0.099	Small	0.0088	Significant
Ensemble for f-test		0.067	Small	0.0381	Significant

Small Cohen's d values of 0.153, 0.123, and 0.029 were noted for SVM, NB, and DT, respectively. Significant gains were seen for SVM (p=0.0014) but not for NB (p=0.2521) or DT (p=0.3540), according to the t-test results. These results highlight the classifier-specific effects of the mRMR feature selection, with SVM gaining greatly from mRMR feature selection—possibly because these characteristics overlap with the decision boundaries of SVM—while NB and DT are barely improved.

Cohen's d values of 0.109 (SVM), 0.216 (NB), and 0.022 (DT) were observed, indicating moderate to medium effect sizes. In line with the effect sizes, the t-test findings showed substantial improvements for NB (p=0.0200) and SVM (p=0.0029). Nevertheless, DT did not show a statistically significant improvement (p=0.5219), despite a small effect size. These findings demonstrate the complex effects of choosing chi-square feature selection model.

NB showed a medium effect size (0.216), with small Cohen's d values of 0.111 (SVM) and -0.031 (DT) noted. The t-test findings suggested substantial improvements for SVM (p=0.0023) and NB (p=0.0299), consistent with the effect sizes. However, despite a small effect size, DT did not show significant improvement (p=0.3262). The fact that the f-test helps SVM and NB but not DT highlights the importance of feature selection compatibility with classifiers.

The application of Cohen's d and t-test for the ensemble models indicated small effect sizes (0.145, 0.099, and 0.067) and significant improvements (p=0.0207, p=0.0088, and p=0.0381) for mRMR, Chi-square test, and f-test, respectively. These findings suggest that the ensemble models, despite their small effect sizes, contributed significantly to the overall improvement, highlighting the complementary nature of selected features when combined through ensemble strategies.

8.3.3 Mod_AlexNet

In Scenario 3, features from the "fc7" layer of Mod_AlexNet were fused with HOG descriptors. Afterwards, a number of feature selection models were applied to identify robust features and lower the dimensionality of the feature vector, including mRMR, chi-square test, and f-test. The features that were selected were then fed into three independent classifiers (SVM, NB, and DT). By integrating the results of

several classifiers using a voting ensemble classification model, the final prediction was computed. Tables 8.5 and 8.6 provide a full breakdown of the evaluation results and the performance indicators that were calculated for the third scenario.

Table 8.5 The performance of Mod_AlexNet-based developed system with different integrated subphases.

Different integrated contexts	Classifier	Performance Measures				
		Accuracy	Sensitivity	Precision	Specificity	F1-Score
Mod_AlexNet only	SVM	93.01%	93.01%	91.45%	29.88%	91.49%
	NB	93.67%	93.67%	92.88%	30.95%	91.98%
	DT	92.08%	92.08%	90.28%	32.48%	90.87%
Mod_AlexNet and Feature Fusion	SVM	93.67%	93.67%	92.60%	34.48%	92.26%
	NB	91.65%	91.65%	90.31%	33.80%	90.74%
	DT	92.58%	92.58%	91.12%	35.50%	91.46%
Mod_AlexNet, Feature Fusion, and mRMR	SVM	94.27%	94.27%	93.51%	40.42%	93.13%
	NB	94.58%	94.58%	94.36%	38.54%	93.35%
	DT	93.34%	93.34%	92.12%	41.56%	92.41%
Mod_AlexNet, Feature Fusion, and chi-square test	SVM	91.86%	91.86%	90.63%	37.23%	91.13%
	NB	93.91%	93.91%	93.01%	37.53%	92.63%
	DT	92.33%	92.33%	90.96%	38.50%	91.46%
Mod_AlexNet, Feature Fusion, and f-test	SVM	93.43%	93.43%	92.47%	37.48%	92.28%
	NB	92.21%	92.21%	91.07%	36.99%	91.36%
	DT	92.51%	92.51%	91.78%	39.88%	91.74%
Ensemble for mRMR		94.91%	94.91%	94.90%	43.07%	93.79%
Ensemble for chi-square test		94.27%	94.27%	94.20%	39.95%	93.10%
Ensemble for f-test		93.63%	93.63%	92.39%	39.86%	92.35%

Using the features of Mod_AlexNet, the performance evaluation of classifiers in many integrated contexts offers an extensive understanding of how feature fusion, selection, and ensemble strategies affect various performance metrics and are provided in Table 8.5. SVM demonstrates that it can classify the features of Mod_AlexNet with a 93.01% accuracy rate. The specificity, at 29.88%, is a little low, though, indicating that it would be difficult to correctly identify negative classes. With a 91.49% F1-Score, recall and precision are well-balanced. Switching to the NB classifier, the overall accuracy is 93.67%, which shows a marginally better capacity for accurate instance classification. The sensitivity and precision metrics, at 93.67% and 92.88%, respectively, are excellent. Additionally, compared to SVM, the NB classifier has a higher specificity of 30.95%, indicating a stronger capacity to

recognise negative cases. With an accuracy of 92.08%, the DT classifier performs competitively. Metrics such as sensitivity, precision, and F1-Score are all above 90%, demonstrating the ability of the model to accurately categorise cases while striking a solid balance between recall and precision. Although the specificity of 32.48% is higher than that of SVM, it is still very low, suggesting that there may be opportunity for improvement in accurately recognising negative cases. To summarise, the first classification scenario with the features of Mod_AlexNet shows good performance for SVM, NB, and DT classifiers. SVM leads in accuracy, NB has marginally better specificity, and DT maintains a good balance across performance metrics.

In the context of classifying the features of Mod_AlexNet fused with HOG descriptors, SVM achieves an impressive 93.67% accuracy, demonstrating a high percentage of correct classifications. Impressive metrics of 93.67% and 92.60%, respectively, are also found in the sensitivity and precision measurements. This implies that the SVM classifier minimises FPs while efficiently identifying cases that are positive. The specificity, at 34.48%, is still somewhat poor, suggesting that it could be difficult to accurately identify negative examples. The overall accuracy drops to 91.65% when comparing this to the NB classifier in the same integrated context. Metrics for precision and sensitivity are still high, at 90.31% and 91.65%, respectively. Even though it has slightly increased when compared to SVM, the specificity—33.80%—remains low. At 90.74%, the F1-Score exhibits a commendable trade-off between precision and recall, which is balanced. When it comes to accurately identifying instances, NB performs well even though its accuracy is lower than that of SVM. In the fused features scenario, the DT classifier exhibits robust performance as its accuracy rises to 92.58%. Metrics such as sensitivity, precision, and F1-Score are all above 90%, demonstrating the ability of the model to accurately categorise cases while maintaining a balance between recall and precision. With a specificity of 35.50%, it is better than both SVM and NB, indicating a higher capacity to recognise negative cases. To summarise, the Mod_AlexNet features fused with HOG descriptors are well-classified by SVM, NB, and DT classifiers, whereas SVM outperforms the other classifiers in terms of all performance measures.

In the integrated context, where the features of Mod_AlexNet are fused with HOG descriptors and then selected using the mRMR feature selection model, the SVM classifier achieves an incredible 94.27% accuracy. Given its high accuracy, SVM appears to be an excellent method for classifying occurrences into the appropriate categories. At 94.27%, 93.51%, and 93.13%, respectively, the sensitivity, accuracy, and F1-Score measures are also quite high. Nevertheless, the specificity metric is comparatively lower at 40.42%, suggesting that it may be difficult to accurately detect negative cases. In contrast, with an accuracy of 94.58% in the same integrated environment, the NB classifier performs remarkably well. Impressively high measures include precision (94.36%), sensitivity (38.54%), and F1-Score (93.35%). Even though the specificity is 38.54%, which is slightly less than SVM, it still shows a good capacity to recognise negative cases. In the scenario with fused and mRMR-selected features, the DT classifier maintains a high accuracy of 93.34%. With values above 92%, the sensitivity, precision, and F1-Score metrics are likewise excellent and demonstrate the ability of the model to accurately identify cases. With a specificity score of 41.56%, it is better than both SVM and NB, suggesting a higher capacity to accurately detect negative cases. In conclusion, all classifiers perform efficiently when the features of Mod_AlexNet are merged with HOG descriptors and selected using the mRMR feature selection model. Outstanding accuracy, precision, and F1-Score metrics are displayed by SVM and NB, demonstrating their effectiveness in classification tasks. Despite having a marginally lower accuracy rate, the DT classifier offers competitive overall performance and demonstrates how feature fusion and selection can improve classification results.

In the integrated context, the SVM classifier achieves 91.86% accuracy when its features are fused with HOG descriptors and subsequently selected using the chi-square feature selection technique. This implies a high degree of efficiency in accurately classifying cases. Significant measures include the F1-Score (91.13%), sensitivity (91.86%), and precision (90.63%). The specificity score, at 37.23%, is comparatively lower, indicating difficulties in accurately recognising negative cases. In comparison, the NB classifier performs remarkably well in the same integrated context, with an accuracy of 93.91%. Impressively high measures are the F1-Score, sensitivity, and precision, which are 37.53%, 92.63%, and 93.01%, respectively. The

specificity, at 37.53%, is marginally lower than SVM, but it still shows a good ability to recognise negative cases. Finally, the accuracy of the DT classifier of 92.33% is still rather high. Notable metrics include sensitivity, precision, and F1-Score, all of which have values above 90% and indicate that the model is effective at properly categorising cases. At 38.50%, the specificity score shows a higher capacity to accurately identify negative instances than SVM, which is an improvement.

The SVM classifier exhibits an accuracy of 93.43% when it uses the f-test feature selection technique using features that have been fused and selected. Sensitivity and Precision, which stand at 93.43% and 92.47%, respectively, indicate how well the model can detect positive cases and how accurate positive predictions are. But the Specificity, which measures how well the model can detect negative examples, is rather low at 37.48%, indicating that this area could use some improvements. The accuracy of the NB classifier is 92.21%, which shows that its predictions are highly accurate. The model demonstrates its capacity to precisely identify positive instances and the accuracy of its positive predictions with sensitivity and precision of 92.21% and 91.07%, respectively. The specificity measurement, is 36.99%, indicating that improvement is needed in the classification of negative cases. The accuracy, sensitivity, and precision of the DT classifier are 92.51%, 91.78%, and 92.51%, respectively. With a specificity measure of 39.88%, it performs better than both SVM and NB at appropriately detecting negative cases. When compared, the SVM classifier performs exceptionally well due to its high accuracy and F1-Score. Even though the NB classifier is less accurate than the SVM, it performs competitively, especially in sensitivity and precision. When compared to the other classifiers, the DT classifier provides a noticeable improvement in specificity.

Several integrated scenarios utilising a voting ensemble classification model were explored, concentrating on features that were fused, chosen using feature selection models such as mRMR, chi-square, and f-test, and then fed into the ensemble classifier. The classifiers attained a noteworthy accuracy of 94.91% when features merged and selected using the mRMR feature selection model were fed into the voting ensemble classification model. At 94.91%, both sensitivity and precision show how well the model can detect positive cases and how accurate its positive predictions are. The accuracy of identifying negative instances, or specificity, is

43.07%, indicating that there is still space for development in this regard. The accuracy of features fused selected with aid of the chi-square feature selection model, which are then fed into the voting ensemble classification model, is 94.27%, indicating a high degree of prediction accuracy in every case. Both sensitivity and precision, at 94.27%, show that positive examples can be identified with effectiveness and positive predictions can be made with precision. The accuracy of detecting negative instances is measured by the Specificity metric, which stands at 39.95%. This indicates that there may be room for improvement in accurately identifying negative occurrences. For the case where features were fused and selected using the f-test feature selection model, and then fed into the voting ensemble classification model, the accuracy is 93.63%, indicating that overall, the predictions were correct; the sensitivity and precision metrics, also 93.63%, show that positive instances were correctly identified, and the specificity metric, 39.86%, indicates that there is room for improvement in correctly identifying negative instances. In comparison, the mRMR-based ensemble model has the highest specificity at 43.07%, indicating that it is adept at correctly recognising negative events. Nonetheless, the ensemble model based on chi-squares exhibits a balanced F1-Score and a little reduced Specificity, signifying its strong overall performance. The ensemble model based on the f-test exhibits competitive metrics, highlighting its efficacy in various integrated situations.

An in-depth understanding of how feature fusion, selection, and ensemble techniques affect different classifiers is provided by the evaluation of improvement in classification performance rates as compared to the limited use of the features Mod_AlexNet for classification. This summary provides insight into the consequences of using different strategies on the overall classification performance, based on the findings shown in Table 8.5. When categorising the features of Mod_AlexNet fused with HOG descriptors, the SVM classifier shows a minor improvement of 0.71% in terms of accuracy. This suggests a marginal improvement in the total number of correctly predicted cases. Conversely, the DT classifier exhibits a positive improvement of 0.55%, whilst the NB classifier shows a noteworthy decline of -2.16%, indicating a reduction in overall accuracy. These differences highlight the complex effects of feature fusion on several classifiers. SVM and DT both demonstrate positive gains in sensitivity, of 0.71% and 0.55%,

respectively, indicating a marginal rise in the accurate detection of positive cases. On the other hand, NB sees a decline of -2.16%, suggesting a drop in sensitivity. This indicates that feature fusion can affect the ability of classifiers to accurately detect positive occurrences, with NB showing a decrease in sensitivity and SVM and DT showing positive impact. The accuracy in identifying positive occurrences is demonstrated by the precision findings; SVM achieved an increase of 1.25%, demonstrating increased precision. In contrast, negative rates of -2.76% and 0.93% for NB and DT, respectively, indicate a decline in precision. The specificity metric exhibits a variety of patterns, indicating the capacity of the classifier to accurately recognise negative instances. SVM shows a notable 15.39% gain, demonstrating better specificity. On the other hand, the positive rates for NB and DT are lower, at 9.21% and 9.32%, respectively.

When the features of Mod_AlexNet are fused with HOG descriptors and selected using the mRMR feature selection model, SVM shows a positive improvement rate of 1.36% in accuracy. This represents a slight improvement in the total accuracy of the categorisation predictions. By comparison, the improvement rates for NB and DT are marginally higher at 0.96% and 1.37%, respectively. Slight variations are revealed by precision, the positive improvement rate for SVM is 2.25%, although the positive rates for NB and DT are less, at 1.59% and 2.04%, respectively. Specificity yielded negative improvement rate of -1.23% for SVM raises the possibility of a specificity compromise. Conversely, NB and DT show somewhat higher positive rates, suggesting marginal gains in accurately identifying negative cases.

The feature selection models f-test and chi-square are explored next. The chi-square feature selection model results in a negative improvement rate in accuracy (-1.23%) and F1-Score (-0.39%) for SVM when HOG descriptors are fused and selected. On the other hand, NB and DT show positive rates for all metrics. SVM shows positive rates of improvement in accuracy (0.45%), sensitivity (0.45%), precision (1.12%), and F1-Score (0.87%) when applied to the f-test feature selection model. This suggests that there has been a positive effect on categorisation performance overall. On the other hand, the majority of measures show negative rates for NB, which suggests that the f-test feature selection technique may have brought trade-offs.

Positive improvement rates are shown by DT, indicating an improvement in F1-Score and precision.

The ensemble model that integrates features fused and selected by the mRMR feature selection approach shows the highest improvement rate, at 3.08%, when it comes to accuracy. This indicates a significant improvement in overall classification accuracy as compared to the use of the features of Mod_AlexNet exclusively. Although showing positive improvement rates of 2.39% and 1.69%, respectively, the chi-square and f-test feature selection models show somewhat smaller accuracy gains. Intriguing patterns can be observed in precision, where the ensemble model utilising characteristics selected by mRMR has the highest precision increase rate, at 5.12%, indicating a notable enhancement in the accuracy of positive predictions. Although they demonstrate positive improvement rates of 2.33% and 4.35%, respectively, chi-square and f-test feature selection models show slightly lower precision gains. The capacity to accurately detect negative instances, or specificity, shows significant improvement rates in all three ensemble models. With a voting ensemble, the mRMR feature selection model obtains the highest improvement rate of 32.63%, demonstrating a significant improvement in accurately recognising negative events. Specificity is also positively impacted by the f-test and chi-square feature selection models, which demonstrate improvement rates of 22.74% and 23.02%, respectively. To sum up, the use of ensemble models that incorporate several feature selection procedures offers a potential way to enhance the performance of image classification. In particular, the mRMR feature selection model shows improved rates across multiple metrics when paired with a voting ensemble classifier. Positive contributions are also made by the f-test and chi-square feature selection models, while their improvement rates are a little bit lower. These findings highlight the significance of choosing suitable feature selection models to achieve desired increases in accuracy, sensitivity, precision, specificity, and F1-Score and offer insightful information.

The t-test and Cohen's d measure were utilised in the assessment as metrics to evaluate the significance and magnitude of the improvements. In this investigation, the performance significance of the classifiers is compared to a baseline that solely uses Mod_AlexNet features and excludes feature fusion, selection, and ensemble

classification models. Table 8.6 displays the results of the Cohen's d measure and the t-test for each context.

Table 8.6 presents a comprehensive and detailed examination of the investigation into several integrated contexts for classifying the features of Mod_AlexNet using Cohen's d measure and t-test for statistical significance. To find out how feature fusion, selection, and ensemble methods affect classification performance, this investigation uses a range of classifiers and feature selection methodologies.

Table 8.6 Comparative Statistical Analysis using Cohen's d and T-test Significance for Evaluating Significance of Improvement against the Performance of Classifiers when Classifying Features Extracted from Mod_AlexNet.

Improvement compared to when classifying features extracted from Mod_AlexNet only	Classifier	Cohen's d:		T-test (p)	
		Measure	Significance	Measure	Significance
Mod_AlexNet and Feature Fusion	SVM	0.066	Small	0.0367	Significant
	NB	-0.037	Small	0.3677	Not Significant
	DT	0.054	Small	0.1079	Not Significant
Mod_AlexNet, Feature Fusion, and mRMR	SVM	0.131	Small	0.1147	Not Significant
	NB	0.093	Small	0.1301	Not Significant
	DT	0.122	Small	0.1636	Not Significant
Mod_AlexNet, Feature Fusion, and chi-square test	SVM	0.038	Small	0.5457	Not Significant
	NB	0.059	Small	0.2800	Not Significant
	DT	0.064	Small	0.3058	Not Significant
Mod_AlexNet, Feature Fusion, and f-test	SVM	0.078	Small	0.2149	Not Significant
	NB	0.005	Small	0.9306	Not Significant
	DT	0.104	Small	0.1293	Not Significant
Ensemble for mRMR		0.201	Medium	0.0321	Significant
Ensemble for chi-square test		0.142	Small	0.0367	Significant
Ensemble for f-test		0.119	Small	0.0767	Not Significant

The SVM classifier exhibited a small Cohen's d of 0.066, indicating a modest effect size, when the features of Mod_AlexNet were fused with HOG descriptors. A significant improvement in performance was indicated by the accompanying t-test, which had a p-value of 0.0367. The NB classifier, on the other hand, showed a non-significant p-value of 0.3677 along with a modest negative Cohen's d of -0.037.

Likewise, a non-significant p-value of 0.1079 was shown by the DT classifier, which had a tiny Cohen's d of 0.054. These findings imply that whereas feature fusion greatly improved the SVM classifier, it had no noticeable impact on NB or DT.

Remarkable discoveries were made when feature selection models were further examined following fusion. When features were chosen using mRMR following fusion, non-significant p-values (0.1147, 0.1301, and 0.1636, respectively) were followed with small Cohen's d values for SVM (0.131), NB (0.093), and DT (0.122). This implies that the improvement in classifier performance in this situation was not that significant in this context.

Small Cohen's d values (0.038, 0.059, and 0.064 for SVM, NB, and DT, respectively) with non-significant p-values (0.5457, 0.2800, and 0.3058) were obtained by applying chi-square feature selection following fusion. Similar to this, small Cohen's d values (0.078, 0.005, and 0.104 for SVM, NB, and DT, respectively) with non-significant p-values (0.2149, 0.9306, and 0.1293) were obtained from f-test feature selection following fusion. These results suggest that performance improvements following fusion were not significantly influenced by either the chi-square or the f-test feature selection models.

When it comes to ensemble techniques, a voting ensemble classification model combined with mRMR-selected features showed a significant p-value of 0.0321 and a medium Cohen's d of 0.201, indicating a significant improvement in performance. Similarly, applying the ensemble model with features selected with the chi-square test produced a significant p-value of 0.0367 and a tiny Cohen's d of 0.142. The ensemble whose features were selected based on the f-test, however, showed a non-significant p-value of 0.0767 together with a small Cohen's d of 0.119, indicating that the f-test had no significant impact on the performance of the ensemble.

Exhibited in Tables 8.1 to 8.6, The performance utilizing Mod_AlexNet is clearly superior to that of SqueezeNet and ResNet-50. The rates of improvement achieved with Mod_AlexNet (Scenario 3) in comparison to SqueezeNet (Scenario 1) and ResNet-50 (Scenario 2) are thus displayed in Table 8.7.

Table 8.7 Comparison of Performance Measures, Cohen's d, and T-test Significance for FFS-EC Implementation Across Different Scenarios, Evaluating Improvement Rates of Scenario 3 Compared to Scenarios 1 and 2

Scenario	Performance Measures					Cohen's d:		T-test (p)	
	Accuracy	Sensitivity	Precision	Specificity	F1-Score	Measure	Significance	Measure	Significance
Scenario 1	2.30%	2.30%	3.63%	21.82%	2.27%	0.1432	Small	0.0003	Significant
Scenario 2	8.92%	8.92%	6.55%	31.96%	6.85%	0.3167	Medium	0.0008	Significant

Table 8.7 compares and thoroughly examines the performance of FFS-EC implementation in two distinct scenarios (Scenario 1 and Scenario 2), with a focus on T-test significance, Cohen's d, and important performance metrics. The main objective is to assess how significantly Scenario 3 improved on Scenarios 1 and 2 in terms of Accuracy, Sensitivity, Precision, Specificity, and F1-Score. In comparison to scenario 1, scenario 3 demonstrates significant improvement of 21.82% in Specificity, but only a moderate improvement of 2.30% in Accuracy, Sensitivity, Precision, and F1-Score. With a Cohen's d value of 0.1432, there is a slight but statistically significant improvement, suggesting a small impact size. The significance of the observed improvements is further highlighted by the T-test significance ($p = 0.0003$), which confirms that scenario 3 performs significantly better than scenario 1. Scenario 3 shows a more noticeable improvement that can be observed across all performance measures in comparison to Scenario 2. Along with a major improvement in specificity of 31.96%, there are significant improvements of 8.92% in the accuracy, sensitivity, precision, and F1-score. In comparison to Scenario 1, the moderate impact size shown by the Cohen's d value of 0.3167 indicates a more noticeable improvement. Strengthening the statistical validity of the observed increases is the T-test significance ($p = 0.0008$). Comparing Scenario 3 against both Scenarios 1 and 2, it is evident that the improvement rates achieved are influenced by the specific characteristics of each scenario. The improvements in accuracy, sensitivity, precision, specificity, and F1-score in Scenario 3 can be attributed to the powerful model of Mod_AlexNet. The effect sizes and statistical significance obtained through Cohen's d and T-test provide a comprehensive understanding of

the practical significance and reliability of the observed improvements in FFS-EC implementation across diverse deep learning models.

A comprehensive comparison of classifier performance revealed notable trade-offs among the models. The SVM classifier demonstrated high sensitivity and accuracy; however, it consistently showed lower specificity, indicating a tendency to misclassify abnormal cases. In contrast, the NB classifier excelled in specificity but at the expense of sensitivity and F1-score, suggesting limitations in detecting positive cases. The DT classifier exhibited balanced performance across most metrics but struggled with specificity in more complex scenarios. These findings underscore the importance of carefully balancing specificity and sensitivity when evaluating model performance, particularly in medical diagnostics where both false positives and false negatives can have significant clinical consequences.

8.4 Summary

In this chapter, the goal of the FFS-EC system was to provide an enhanced framework for the classification of DBT data through the integration of deep learning models with feature fusion and selection models, followed by an ensemble classifier. It addresses three primary challenges: low performance measures in multi-class classification scenarios; the complexity of different breast sizes and densities; and distinguishing between benign and malignant abnormalities.

Work in this chapter relies significantly on deep learning models, with ResNet-50, SqueezeNet, and the previously developed Mod_AlexNet making distinct contributions to the performance of the system. With its deep residual architecture, ResNet-50 effectively reduces the vanishing gradient problem and is an excellent model for capturing complex characteristics. This is especially important when dealing with the challenging task of differentiating between various breast classes because the depth of ResNet-50 enables it to collect feature hierarchies, which helps identify modest abnormalities across a range of densities and sizes of breasts. The compact architecture of the SqueezeNet improves performance without compromising feature extraction capabilities. With its carefully chosen layer modifications, Mod_AlexNet tackles issues like internal covariate shift and overfitting, making learning more accurate, robust, and discriminative. This system also incorporates the HOG and the HSV colour scheme to enhance feature extraction,

enabling enhanced discrimination between benign and malignant tumours by emphasising edges, patterns, and colour information.

Improving discriminating ability is largely dependent on the feature selection and fusion processes. mRMR is employed to measure the relevance and redundancy of fused features. It evaluates the mutual information between each feature and the target class (relevance) while considering the redundancy between features. The chi-square test evaluates the degree of independence between categorical classes in the context of DBT data, highlighting features that have strong correlations with the target class. Similarly, the f-test contributes significantly to the feature selection process. This test helps identify features that significantly contribute to the differentiation of benign and malignant abnormalities by assessing the statistical significance of mean differences across groups. An advanced integration approach called the voting ensemble model, which aggregates predictions from several classifiers, is deployed to get a final classification prediction. By using a voting mechanism, this ensemble technique combines the advantages of various classifiers, such as SVM, NB, and DT. This approach addresses the multi-class classification challenge by leveraging the strengths of several classifiers, improving accuracy, sensitivity, and specificity.

The FFS-EC system demonstrated significant improvements in classification accuracy and sensitivity, particularly through the integration of feature fusion and selection techniques. However, challenges with specificity persist, especially in accurately distinguishing between benign and malignant cases. Overfitting risks were managed through the implementation of cross-validation and feature reduction strategies, ensuring more reliable model performance. The hybrid approach, which combined HOG and deep learning-derived features, effectively balanced the strengths of low- and high-dimensional feature representations, though some trade-offs remain in classifier sensitivity and specificity.

The effectiveness of Mod_AlexNet in this system is thoroughly examined and compared to ResNet-50 and SqueezeNet, showing that Mod_AlexNet significantly outperforms the others, especially when utilizing mRMR feature selection. While Mod_AlexNet provided an important foundation in addressing the challenges of varying breast densities, results proved that the ability to accurately classify

abnormalities was limited. This highlights the need for more sophisticated models, leading to the development of FFS-EC, which incorporated feature fusion and selection models to address these challenges. However, despite the improvement made by the FFS-EC System, multi-class performance, especially in specificity measure, remained a challenge, prompting the development of the next system, presented in Chapter 9, to enhance classification performance.

Chapter 9 Three-Layer Multi-Head Self-Attention Model for Enhanced DBT Classification

9.1 Introduction

Despite notable advancements in specificity, sensitivity, and accuracy presented and discussed in Chapter 8, some challenges are yet to be addressed successfully, prompting the development of a novel three-layer Multi-Head Self-Attention model, as presented in this chapter. While previous systems, such as the FFS-EC System, made significant progress, particularly in improving feature fusion and selection, the performance of correct abnormal classification across varying breast densities and sizes remains a challenge.

To tackle the challenge of varying breast and tumour sizes, a system is introduced in this chapter that integrates a novel self-attention model with a deep learning model to improve the ability of the model to identify abnormalities regardless of changes in tumour size by enabling the model to selectively focus on relevant regions within the breast image. By selecting features that are essential for classification, the attention mechanism helps the model overcome the challenges. Furthermore, the self-attention mechanism improves the robustness of the model and adaptability by dynamically varying the weight assigned to various spatial areas within the breast image, which raises the diagnostic accuracy of tomosynthesis imaging overall.

9.2 Methodology

In this system, a variety of pre-processing procedures were developed to augment and enhance the images. To increase model generalisation and enrich the dataset, augmentation processes were first applied to the images. Pre-processing methods were then used to improve the clarity and quality of the images. Notably, the discriminative qualities of the images were further improved by incorporating the HSV colour mapping technique. Following the colour mapping, the Mod_AlexNet architecture was applied in conjunction with an entirely novel self-attention model designed explicitly to address the varying size in the breast and tumour challenge. The techniques that were deployed in this system for the augmentation, pre-processing, and colour mapping, were explained in depth in Chapter 3.

The third contribution is the development of a three-layer Multi-Head Self-Attention model, Multi-Head Mod_AlexNet Attention (MHMA), using the final pooling outputs of the Mod_AlexNet. This model integrates FC layers and self-attention models, developed specifically to address problems caused by variations in breast sizes in DBT data and the varied sizes of the tumours. Through the use of SGDM, Adam, and RMSProp optimisers during training across a variety of batch sizes, the adaptability of the model is increased. The meticulous procedures employed in MHMA System resulted in a significant improvement in classification performance and will be presented in the following section. The architecture of the developed MHMA System is shown in Figure 9.1.

The dataset used in this study consists of real-world digital breast tomosynthesis (DBT) images, which improves the generalizability and clinical relevance of the results. While augmentation and preprocessing steps enhance the diversity of the training set, careful validation ensured that the augmented images did not introduce artifacts or distort real-world patterns. The input images were augmented, and four images were generated from a single image. The use of Gaussian smoothing and CLAHE enhancement helps maintain the natural characteristics of the breast tissue images while reducing noise and improving contrast. The HSV colour scheme was utilised for colour mapping after the enhancement. The images were fed into the Mod_AlexNet, which was merged with the newly developed self-attention model.

The development of a self-attention model designed for multi-class tomosynthesis classification, with three Multi-Head self-attention layers and the outputs of the final three pooling layers from Mod_AlexNet, is an important advancement in medical image analysis. This approach uses attention mechanisms to extract complex information from tomosynthesis images, resulting in more accurate classification across many classes. A key component of this approach is self-attention, which makes it easier to examine complex connections within the input data and identify patterns in tomosynthesis images that correspond to distinct classes. Through learnt linear projections, the model converts input feature representations into queries, keys, and values. This process computes attention scores, which enable the model to concentrate on significant regions and features within the images.

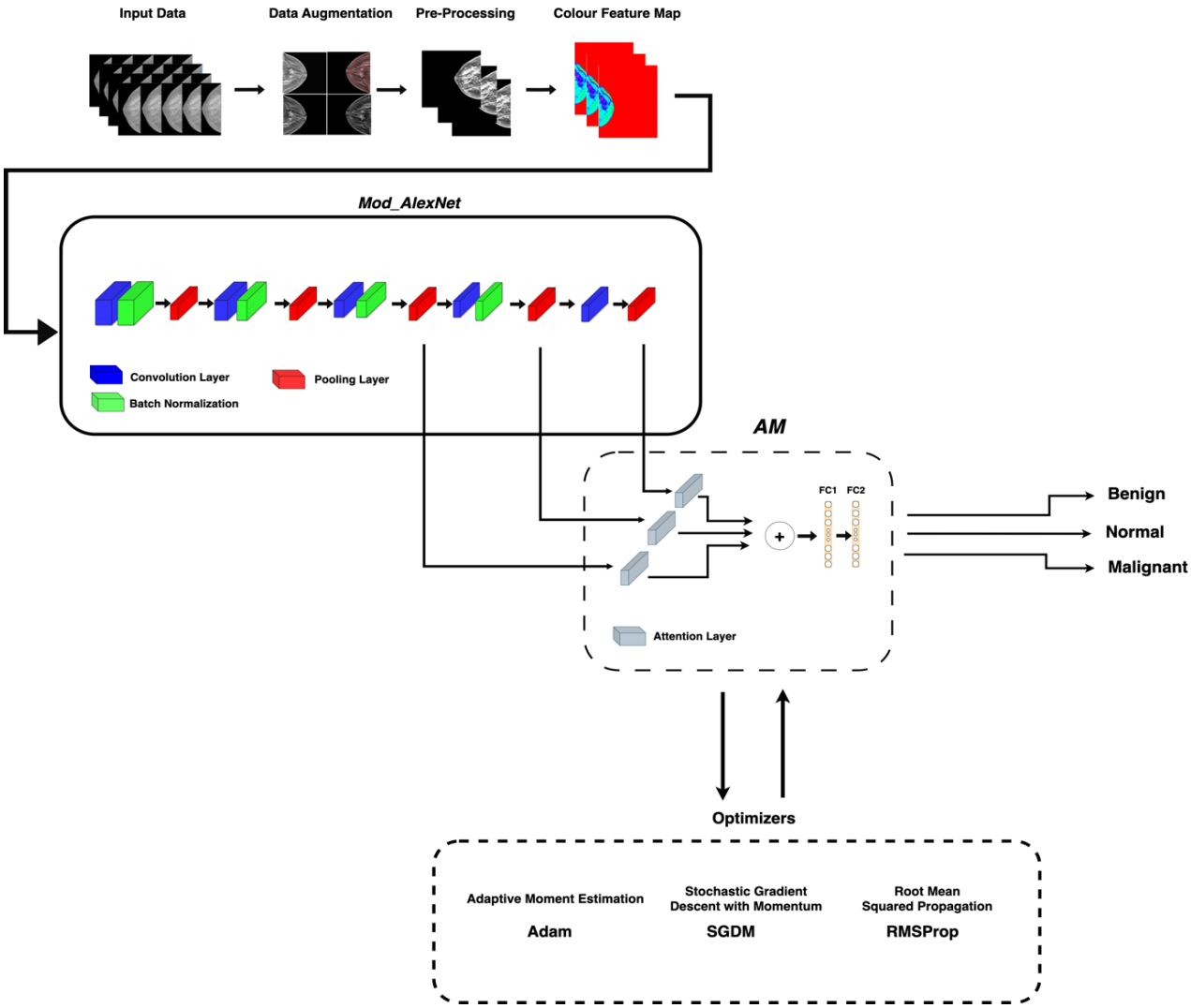

Figure 9.1 Architecture of the Multi-Head Mod_AlexNet Attention (MHMA) System

The initial Multi-Head Self-Attention layer starts the process by detecting various correlations and patterns in the tomosynthesis images. The ability of this layer to extract discriminative features relevant to the classification task is improved by its multiple attention heads, which allow it to attend to different components of the input data simultaneously. The learnt representations are further enhanced and refined by additional Multi-Head Self-Attention layers, which repeatedly capture hierarchical dependencies in the images. Through this iterative process, the model is able to extract features that are more abstract and contextually rich, leading to more accurate classification across a variety of classifications. The rich feature representation that is collected from the tomosynthesis images by the last three pooling layers of Mod_AlexNet gives the self-attention model a complete input on which to base its classification decisions. By utilising these features, the model is

better able to distinguish minute variations between several classes, which enhances classification performance. Compared to a self-attention model with only one attention layer, the addition of several attention layers improves the ability of the model to capture complicated relationships within the input data, allowing it to reliably identify tomosynthesis images across various classes. Each layer refines the representations learnt in the previous levels, increasing the overall discriminative power.

The attention mechanisms and corresponding maps automatically highlight parts of tomosynthesis images that are relevant and distinctive to each of the three classes: normal, benign, and malignant. Several attention mechanisms were used in our system at different levels of pooling. Initially, just two attention models were used, ranging from the initial to the final convolution level. The count of attention mechanisms was then raised to three, and then to four. The integration of three attention mechanisms from the final pooling layers resulted in optimal performance across all metrics, since the deeper layers of the deep learning model provide richer information and are better suited to extracting features that aid in identifying the three classes under investigation.

These three attention processes are used to analyse the intermediate feature maps generated by Pools 3, 4, and 5. The final feature vector is then created by concatenating the resulting feature vectors, which are then concatenated. This final feature vector is used as the input for two FC layers. This attention module was trained using a variety of optimisers, including Adam, RMSProp, and SGDM, in addition to varying batch sizes.

Multi-Head Self-Attention layers play an important role in deep learning architectures for image categorisation because of their capacity to capture complicated linkages and contextual information inside feature maps generated by pooling layers. These layers simultaneously handle different parts of the input feature map, enabling the model to identify significant patterns and connections between features, improving classification resilience and accuracy. Multi-Head Self-Attention layers play a major role in the ability of the model to extract discriminative features and produce well-informed classification judgements in the context of image classification. Figure 9.2 shows the architecture of a Multi-Head Attention module (Vaswani et al., 2017).

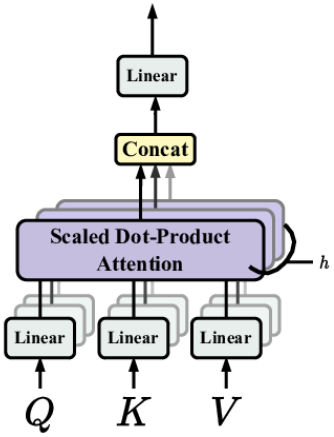

Figure 9.2 Multi-Head Attention Module presented by (Vaswani et al., 2017).

Multi-Head Self-Attention layers operate via a number of learnt parameters and mathematical operations. The layer uses linear transformations to calculate query, key, and value matrices—typically denoted as Q, K, and V, respectively—after receiving feature mappings from pooling layers (Vaswani et al., 2017). These matrices form the basis for evaluating the relative importance of features by computing attention scores between pairs of features. The attention score α_{ij} between features i and j is calculated as follows in equation 1:

$$\alpha_{ij} = softmax\left(\frac{Q_{i*}K_j^T}{\sqrt{d_k}}\right) \qquad (9.1)$$

Where Q_i and K_j represent the query and key vectors derived from features i and j, respectively, and d_k is the dimensionality of the key vectors (Vaswani et al., 2017). A probabilistic definition of attention weights is provided by the softmax function, which guarantees that the attention scores across all features add up to 1. Following that, the attended representation y_i for feature i is calculated as the weighted sum of the value vectors V_j for every feature, which is then weighted by the attention scores α_{ij}.

$$y_i = \sum_{j=1}^{n} \alpha_{ij} * V_j \qquad (9.2)$$

The total number of features in the input feature map is shown by n in this case. Within the Multi-Head Self-Attention layer, each attention head has a unique set of

learnable parameters, such as query, key, and value matrices, which are optimised through backpropagation during training.

Figure 9.3 displays the design of the developed attention model, which represents the third key contribution, and Table 9.1 presents the architecture of the integrated Mod_AlexNet with the layers parameters of the Multi-Head Self-Attention model. Self-attention layers are part of an intentional attempt to improve the capacity of the model to extract complex dependencies and relevant features from the input sequences. This is particularly evident when considering the relationships between the Pool3, Pool4, and Pool 5 levels. With carefully determined parameters, each self-attention layer is set up to maximise representation learning and feature extraction.

Starting with Selfattention_2, which is connected to the Pool3 layer, the layer has 8 heads, 64 key and query channels, 256 value channels, and 256 output channels. The selection of these attributes aims to achieve a balance between representational capacity and computing efficiency. The inclusion of 8 heads enables the network to attend to several sections of the input space at the same time, making it easier to extract diverse and discriminative features. The layer effectively computes attention scores while maintaining sufficient capacity to capture intricate patterns by assigning 64 key and query channels. Moreover, the 256 value channels of the layer allow it to encode deep contextual information, which strengthens feature representations. From Pool 4 onwards, similar parameter selections are noted to preserve consistency with the architectural layout and the features of the previous feature space in the self-attention layer that follows. By using 8 heads, parallelised processing is made possible, improving computational performance without sacrificing representational capacity.

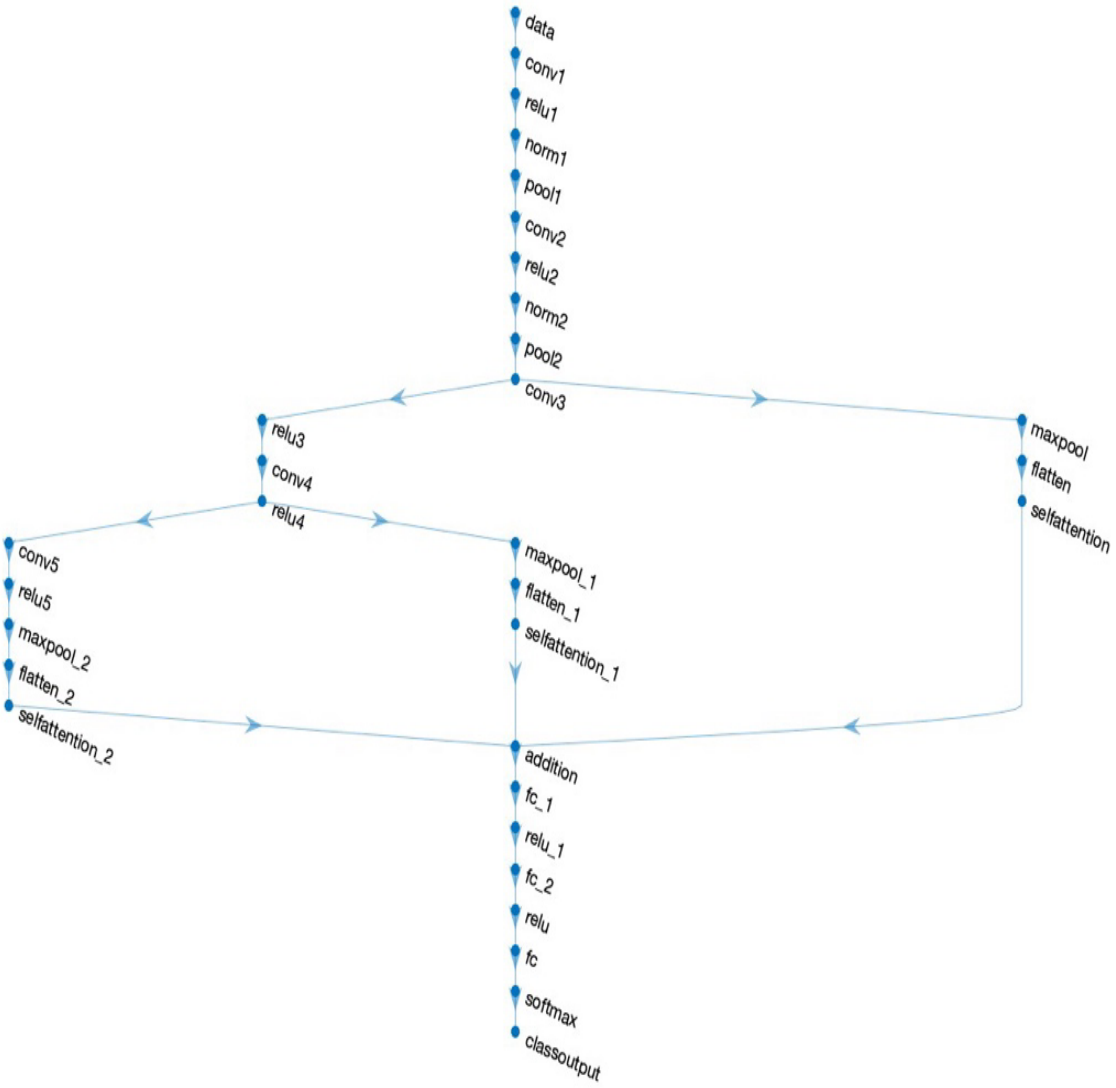

Figure 9.3 The Innovative Multi-Head Attention Model and the Integrated Mod_AlexNet architecture

Table 9.1 Layer Parameters for the Innovative Multi-Head Mod_AlexNet Attention (MHMA) model

Layer Number	Name	Type	Activations	Learnable Properties
1	Data 227x227x3 images with 'zerocenter' normalization	Image Input	227(S) x 227(S) x 3(C) x 1(B)	—
2	Convl 96 11x11 convolutions with stride [4 4] and padding [0 0 0 0]	2-D Convolution	55(S) x 55(S) x 96(C) x 1(B)	Weights 11 x 11 x 3 x 96 Bias 1 x 1 x 96
3	Relul ReLU	ReLU	55(S) x 55(S) x 96(C) x 1(B)	—
4	Norml cross channel normalization with 5 channels per element	Cross Channel Normalization	55(S) x 55(S) x 96(C) x 1(B)	—
5	Pool1 3x3 max pooling with stride [2 2] and padding [0 0 0 0]	2-D Max Pooling	27(S) x 27(S) x 96(C) x 1(B)	—
6	Conv2 2 groups of 128 5x5 convolutions with stride [1 1] and padding [2 2 2 2]	2-D Grouped Convolution	27(S) x 27(S) x 256(C) x 1(B)	Weights 5 x 5 x 48 x 128 x 2 Bias 1 x 1 x 128 x 2
7	Relu2 ReLU	ReLU	27(S) x 27(S) x 256(C) x 1(B)	—
8	Norm2 cross channel normalization with 5 channels per element	Cross Channel Normalization	27(S) x 27(S) x 256(C) x 1(B)	—
9	Pool2 3x3 max pooling with stride [2 2] and padding [0 0 0 0]	2-D Max Pooling	13(5) x 13(5) x 256(C) x 1(B)	—
10	Conv3 384 3x3 convolutions with stride [1 1] and padding [1 1 1 1]	2-D Convolution	13(S) x 13(S) x 384(C) x 1(B)	Weights 3 x 3 x 256 x 384 Bias 1 x 1 x 384
11	Relu3 ReLU	ReLU	13(S) x 13(S) x 384(C) x 1(B)	—
12	Conv4 2 groups of 192 3x3 convolutions with stride [1 1] and padding [1 1 1 1]	2-D Grouped Convolution	13(S) x 13(S) x 384(C) x 1(B)	Weights 3 x 3 x 192 x 192 x 2 Bias 1 x 1 x 192 x 2
13	Relu4 ReLU	ReLU	13(S) x 13(S) x 384(C) x 1(B)	—
14	Conv5 2 groups of 128 3x3 convolutions with stride [1 1] and padding [1 1 1 1]	2-D Grouped Convolution	13(S) x 13(S) x 256(C) x 1(B)	Weights 3 x 3 x 192 x 128 x 2 Bias 1 x 1 x 128 x 2
15	Relu5 ReLU	ReLU	13(S) x 13(S) x 256(C) x 1(B)	—
16	Maxpool_2 7x7 max pooling with stride [1 1] and padding 'same'	2-D Max Pooling	13(5) x 13(S) x 256(C) x 1(B)	—
17	Flatten_2 Flatten	Flatten	43264(C) x 1(B)	—

18	Selfattention_2 Self attention layer with 256 output channels, 8 heads, 64 key and query channels, and 256 value channels	Self Attention	256(C) x 1(B)	QueryWeights 64 x 43264 KeyWeights 64 x 43264 ValueWeights 256 x 43264 OutputWeights 256 x 256 QueryBias 64 x 1 KeyBias 64 x 1 ValueBias 256 x 1 OutputBias 256 x 1
19	Maxpool 7x7 max pooling with stride [1 1] and padding 'same'	2-D Max Pooling	13(S) x 13(S) x 384(C) x 1(B)	—
20	Flatten Flatten	Flatten	64896(C) x 1(B)	—
21	Selfattention Self attention layer with 256 output channels, 8 heads, 64 key and query channels, and 256 value channels	Self Attention	256(C) x 1(B)	QueryWeights 64 x 64896 KeyWeights 64 x 64896 ValueWeights 256 x 64896 OutputWeights 256 x 256 QueryBias 64 x 1 KeyBias 64 x 1 ValueBias 256 x 1 OutputBias 256 x 1
22	Maxpool_1 7x7 max pooling with stride [1 1] and padding 'same'	2-D Max Pooling	13(S) x 13(S) x 384(C) x 1(B)	—
23	Flatten_1 Flatten	Flatten	64896(C) x 1(B)	—
24	Selfattention_1 Self attention layer with 256 output channels, 8 heads, 64 key and query channels, and 256 value channels	Self Attention	256(C) x 1(B)	QueryWeights 64 x 64896 KeyWeights 64 x 64896 ValueWeights 256 x 64896 OutputWeights 256 x 256 QueryBias 64 x 1 KeyBias 64 x 1 ValueBias 256 x 1 OutputBias 256 x 1
25	Addition Element-wise addition of 3 inputs	Addition	256(C) x 1(B)	—
26	Fc_1 1000 fully connected layer	Fully Connected	1000(C) x 1(B)	Weights 1000 x 256 Bias 1000 x 1
27	Relu_1 ReLU	ReLU	1000(C) x 1(B)	—
28	Fc_2 1000 fully connected layer	Fully Connected	1000(C) x 1(B)	Weights 1000 x 1000 Bias 1000 x 1
29	Relu ReLU	ReLU	1000(C) x 1(B)	—
30	Fc 3 fully connected layer	Fully Connected	3 (C) x 1(B)	Weights 3 x 1000 Bias 3 x 1

31	Softmax Softmax	Softmax	3(C) x 1(B)	—
32	classoutput crossentropyex	Classification Output	3 (C) x 1(B)	—

Furthermore, the network is able to encode rich contextual information since 64 key and query channels and 256 value channels are distributed in a way that balances computational complexity and feature richness. Additionally, the Selfattention_1 layer connected to the Pool5 layer displays similar parameters that were selected with care to match the underlying architecture of the network and the characteristics of the input features. The consistency in parameter selection between self-attention layers promotes smooth network architecture integration and compatibility. Standardised parameter setup improves computational efficiency and strengthens the processes of representation learning and feature extraction.

A crucial step in the architecture is the concatenation of the outputs of the three self-attention layers. This allows for the integration of various contextual data and feature representations. The outputs from the three self-attention layers were added with the use of this concatenation operation, called "Addition," which produces a combined feature representation. The combined feature representation is then passed through a sequence of FC layers, each of which is intended to further refine and abstract the features for operations that come after the concatenation operation. The next set of ReLU activation functions adds non-linearity, allowing the network to model complex connections in the data. To convert the feature representations obtained from earlier layers into a format appropriate for the final classification task, the FC layer, designated as "Fc_1" with 1000 neurons, is added to the output layer. This addition is an essential component of the architecture. Every parameter in Fc_1, such as the bias vector and weight matrix, is selected to maximise the ability of the network to learn discriminative features and enable precise classification. Similar to this, the Fc_1 bias vector has 1000x1 dimensions, giving every output layer neuron an extra degree of freedom to control how it is activated. The bias terms allow the network to make shifts or offsets in the activation functions, allowing it to learn more flexible decision limits and adapt to changing input distributions. The network becomes more flexible in recognising subtle patterns and maximising its performance on the

classification task by introducing biases into the computation. The last layer classifies images into three groups: benign, malignant, or normal. Findings for MHMA System will be outlined and examined in the subsequent subsection.

9.3 Results and Discussion

In this study, a Multi-Head Mod_AlexNet Attention (MHMA) system was assessed, concentrating on its third contribution, the Multi-Head attention model. Training was carried out on this model using different optimisers and batch sizes. The next section provides a thorough examination of the results obtained from the deployment of MHMA System. In order to determine the efficacy of the Multi-Head Attention model integration with Mod_AlexNet, a thorough assessment was conducted that included a number of performance indicators, such as F1 score, sensitivity, specificity, precision, and accuracy.

The MHMA model was trained and evaluated on a real-world DBT dataset, ensuring that the model's performance reflects real clinical scenarios. The model's ability to generalize was confirmed through 10-fold cross-validation, which showed consistent accuracy and sensitivity across different splits of the dataset. The strong performance across independent validation sets suggests that the model is capable of handling the natural variability present in clinical breast images.

By comparing the performance of the MHMA System to that of the MA System, where Mod_AlexNet was used without the Attention Model integrated, the statistical significance of any observed gains was assessed using Cohen's d and a t-test. Additionally, the investigation involved comparing the output of the MHMA System to the most promising outcomes attained by MA System, with the goal of estimating the importance of the improvements noted. Tables 9.2 through 9.10 show the final outcomes of the developed Multi-Head Attention model post-integration with Mod_AlexNet involving SGDM, RMSProp, and Adam optimizers, considering batch sizes of 32, 62, and 128. The first column in every table indicates the Attention Model that was applied. The layers from which features were transferred to the Attention Model are denoted by the numbers that follow, for example, M1 for the Attention Module, which consists of two Multi-Head Attention layers and M2 for the Attention Module, which consists of three Multi-Head Attention layers. For example, M1_Pool 2,3 indicates that features were transferred from the second and third

pooling levels to the two attention layers in order to build Attention Model M1, and then their outputs and inputs were concatenated into the FC layers. On the other hand, M2_Pool2,3,4 denotes the transfer of features to the three Attention layers from the second, third, and fourth pooling layers. This produces Attention Model M2, which is then combined with its input and output into the FC layers.

9.3.1 Batch Size 32

Table 9.2 presents the performance metrics of MHMA System on Subset 3, which involves merging the Attention Model and Mod_AlexNet with the SGDM optimiser. The learning rate of 0.0001, 50 epochs, and 32 batch sizes were employed. Each row represents a different configuration of the Attention Model, specifically the pooling layers from which features were transferred. Examining the M1 setups, it can be seen that as the model moves towards deeper pooling layers, all performance indicators consistently improve. A sensitivity/recall of 85.35%, accuracy of 85.35%, precision of 88.24%, specificity of 37.40%, and F1-score of 86.68% are obtained from M1_Pool 1,2. These measures increase to 85.84%, 85.84%, 88.49%, 39.24%, and 87.05%, respectively, as the model proceeds to M1_Pool 4,5. As demonstrated by the greater precision and specificity values, this pattern shows that incorporating information from deeper layers improves the ability of the model to correctly classify cases and reduce FPs.

Table 9.2 *Performance metrics evaluated for MHMA System on Subset 3 integrating the Attention Model with Mod_AlexNet utilizing the SGDM optimizer at a learning rate of 0.0001, 50 epochs, and batch size of 32.*

Model/FC Layer	Accuracy	Sensitivity/Recall	Precision	Specificity	F1-Score
M1_Pool 1,2	85.35%	85.35%	88.24%	37.40%	86.68%
M1_Pool 2,3	85.45%	85.45%	88.35%	38.92%	86.78%
M1_Pool 3,4	85.58%	85.58%	88.44%	39.04%	86.88%
M1_Pool 4,5	85.84%	85.84%	88.49%	39.24%	87.05%
M2_Pool 1,2,3	85.80%	85.80%	88.38%	39.05%	86.99%
M2_Pool 2,3,4	86.27%	86.27%	88.17%	39.55%	87.18%
M2_Pool 3,4,5	86.46%	86.46%	88.21%	39.89%	87.30%

Comparably, when investigating M2 deployments, a similar pattern of performance metrics enhancing as pooling layers get deeper. In comparison, M2_Pool 3,4,5 achieves higher values across all metrics, with an accuracy of 86.46%,

sensitivity/recall of 86.46%, precision of 88.21%, specificity of 39.89%, and an F1-score of 87.30%. In contrast, M2_Pool 1,2,3 exhibits an accuracy of 85.80%, sensitivity/recall of 85.80%, precision of 88.38%, specificity of 39.05%, and an F1-score of 86.99%. This emphasises how crucial it is to consider a larger variety of features from deeper layers to enhance model performance.

Table 9.3 *Performance metrics evaluated for MHMA System on Subset 3 integrating the Attention Model with Mod_AlexNet utilizing the Adam optimizer at a learning rate of 0.0001, 50 epochs, and batch size of 32.*

Model/FC Layer	Accuracy	Sensitivity/Recall	Precision	Specificity	F1-Score
M1_Pool 1,2	84.06%	84.06%	87.48%	37.00%	85.70%
M1_Pool 2,3	84.74%	84.74%	87.79%	37.60%	86.20%
M1_Pool 3,4	85.39%	85.39%	88.00%	38.21%	86.64%
M1_Pool 4,5	86.17%	86.17%	88.12%	38.36%	87.11%
M2_Pool 1,2,3	85.89%	85.89%	88.04%	37.61%	86.93%
M2_Pool 2,3,4	85.94%	85.94%	88.47%	38.22%	87.10%
M2_Pool 3,4,5	86.09%	86.09%	88.54%	38.37%	87.12%

Table 9.3 displays the performance metrics assessed for MHMA System on Subset 3, where the Attention Model is combined with Mod_AlexNet through the use of the Adam optimiser. A batch size of 32 is employed, and a learning rate of 0.0001 is applied over a period of 50 epochs. Examining the M1 configurations, it is proven that most performance indicators improve as the model goes deeper into the pooling layers. The results of M1_Pool 1,2 are as follows: 84.06% accuracy, 84.06% sensitivity/recall, 87.48% precision, 37.00% specificity, and 85.70% F1-score. These measures increase to 86.17%, 86.17%, 88.12%, 38.36%, and 87.11%, respectively, as the model proceeds to M1_Pool 4,5. This pattern suggests that incorporating information from deeper layers improves overall performance measures, especially precision and specificity, and improves the ability of the model to identify abnormal instances reliably.

Comparably, while looking at the M2 configurations, a comparable pattern of performance measures getting better with deeper pooling layers can be seen. In comparison, M2_Pool 3,4,5 achieves higher values across all metrics, with an accuracy of 86.09%, sensitivity/recall of 86.09%, precision of 88.54%, specificity of 38.37%, and an F1-score of 87.12%. For example, M2_Pool 1,2,3 exhibits an

accuracy of 85.89%, sensitivity/recall of 85.89%, precision of 88.04%, specificity of 37.61%, and an F1-score of 86.93%. Overall, the findings emphasise the importance of employing multi-layered feature representations to improve the predictive capabilities of the system, with deeper layers contributing to enhanced performance across various evaluation measures.

Table 9.4 Performance metrics evaluated for MHMA System on Subset 3 integrating the Attention Model with Mod_AlexNet utilizing the RMSProp optimizer at a learning rate of 0.0001, 50 epochs, and batch size of 32.

Model/FC Layer	Accuracy	Sensitivity/Recall	Precision	Specificity	F1-Score
M1_Pool 1,2	83.94%	83.94%	86.16%	25.24%	85.03%
M1_Pool 2,3	84.19%	84.19%	86.17%	25.13%	85.16%
M1_Pool 3,4	84.29%	84.29%	86.14%	26.01%	85.20%
M1_Pool 4,5	84.35%	84.35%	86.30%	26.03%	85.31%
M2_Pool 1,2,3	84.31%	84.31%	86.14%	26.04%	85.21%
M2_Pool 2,3,4	84.38%	84.38%	86.31%	26.03%	85.32%
M2_Pool 3,4,5	84.40%	84.40%	86.31%	26.03%	85.34%

The performance metrics assessed for MHMA System on Subset 3 are displayed in Table 9.4. The system integrates the Attention Model and Mod_AlexNet using the RMSProp optimiser, employing a batch size of 32 and a learning rate of 0.0001 across 50 epochs. Every row in the table shows a different attention model setup, indicating the pooling layers that are used to extract features. Analysing the M1 configurations, most performance indicators improve steadily as the model moves towards deeper pooling layers. M1_Pool 1,2, for example, obtains 83.94% accuracy, 83.94% sensitivity/recall, 86.16% precision, 25.24% specificity, and an F1-score of 85.03%. With M1_Pool 2,3, there is a slight improvement as the accuracy, sensitivity/recall, precision, specificity, and F1-score rise to 84.19%, 86.17%, and 85.16%, respectively. While for M1_Pool 4,5, there is an improvement across all metrics and achieving a 84.35% accuracy, 26.03% specificity and an F1-score of 85.31%. Analysing the M2 configurations also demonstrates a tendency towards parallel performance improvement with deeper pooling layers.

M2_Pool 1,2,3, for example, obtains 84.31% accuracy, 84.31% sensitivity/recall, 86.14% precision, 26.04% specificity, and an F1-score of 85.21%. With M2_Pool 2,3,4, slight enhancements can be observed in the majority of metrics: the F1-score

increases to 85.32%, accuracy to 84.38%, sensitivity/recall to 84.38%, precision to 86.31%, and specificity to 26.03%. With an accuracy of 84.40%, precision of 86.31%, and an F1-score of 85.34%, M2_Pool 3,4,5 ultimately demonstrated the optimal performance.

Examining the results from Tables 9.2, 9.3, and 9.4 provides interesting new insights into how MHMA System operates on Subset 3 with different optimiser configurations. Each table displays the integration of the Mod_AlexNet and Attention Model with the different optimisers, SGDM, Adam, and RMSProp, respectively. Overall, M2_Pool 3,4,5 consistently outperforms all three optimiser setups, achieving the greatest performance of any configuration analysed in each table. The SGDM optimiser demonstrated the highest performance among all the optimisers. M2_Pool 3,4,5 yielded the optimum results, with an accuracy of 86.46%, a specificity of 39.89%, and an F1-score of 87.30%. In comparison to other configurations and different optimisers, this configuration demonstrates the highest results across all performance metrics.

9.3.2 Batch Size 64

Tables 9.5, 9.6, and 9.7 investigate and analyse the outcomes obtained from training MHMA System with a learning rate of 0.0001 over 50 epochs and a batch size of 64, employing SGDM, Adam, and RMSProp optimisers, respectively.

Table 9.5 displays the performance metrics that were assessed for MHMA System, with different Attention Model and Mod_AlexNet configurations used during training. A learning rate of 0.0001 was used, along with 50 epochs and a batch size of 64. Every entry in the table represents a specific setup indicated by the pooling layers that are used to extract features. As the model move towards deeper pooling layers, a constant improvement can be seen in most performance metrics, starting with the M1 configurations. M1_Pool 1,2, for example, obtains 89.54% accuracy, 89.54% sensitivity/recall, 90.49% precision, 44.42% specificity, and an F1-score of 89.80%. A minor improvement in these parameters is seen with M1_Pool 2,3, where the F1-score rises to 89.74%, accuracy to 89.65%, sensitivity/recall to 89.65%, precision to 90.23%, and specificity to 44.84%. After that, as the model moves on to M1_Pool 4,5, the metrics continue to get better: 90.63% accuracy, 90.63% sensitivity/recall, 91.64% precision, 50.47% specificity, and a 90.92% F1-score imply that

incorporating features from deeper layers improves the overall performance of the model metrics and its ability to classify abnormal instances accurately.

Table 9.5 Performance metrics evaluated for MHMA System on Subset 3 integrating the Attention Model with Mod_AlexNet utilizing the SGDM optimizer at a learning rate of 0.0001, 50 epochs, and batch size of 64.

Model/FC Layer	Accuracy	Sensitivity/Recall	Precision	Specificity	F1-Score
M1_Pool 1,2	89.54%	89.54%	90.49%	44.42%	89.80%
M1_Pool 2,3	89.65%	89.65%	90.23%	44.84%	89.74%
M1_Pool 3,4	90.51%	90.51%	91.27%	48.36%	90.64%
M1_Pool 4,5	90.63%	90.63%	91.64%	50.47%	90.92%
M2_Pool 1,2,3	90.56%	90.56%	91.57%	50.47%	90.81%
M2_Pool 2,3,4	90.76%	90.76%	91.68%	50.48%	90.98%
M2_Pool 3,4,5	90.99%	90.99%	91.98%	51.99%	91.26%

Analysing the M2 configurations also shows that deeper pooling layers tend to improve performance in a similar way. M2_Pool 1,2,3, for example, obtains 90.56% accuracy, 90.56% sensitivity/recall, 91.57% precision, 50.47% specificity, and 90.81% F1-score. Moving on to M2_Pool 3,4,5, it can be seen that the greatest values of all the metrics: 90.99% accuracy, 90.99% sensitivity/recall, 91.98% precision, 51.99% specificity, and 91.26% F1-score. These results highlight how crucial it is to take into account a wider variety of variables from deeper layers in order to enhance model performance.

The performance results for MHMA System on Subset 3 are shown in Table 9.6. This system integrates the Attention Model and Mod_AlexNet with the Adam optimiser, using a batch size of 64 and a learning rate of 0.0001 across 50 epochs. When the M1 setups are analysed, deeper pooling layers result in better configurations. M1_Pool 1,2, for instance, obtains 86.45% accuracy, 86.45% sensitivity/recall, 88.21% precision, 39.89% specificity, and 87.29% F1-score. With M1_Pool 4,5, these metrics demonstrate minor gains, with sensitivity/recall, precision, specificity, and F1-score reaching 86.49%, 88.26%, 40.08% and 87.34%, respectively.

Table 9.6 Performance metrics evaluated for MHMA System on Subset 3 integrating the Attention Model with Mod_AlexNet utilizing the Adam optimizer at a learning rate of 0.0001, 50 epochs, and batch size of 64.

Model/FC Layer	Accuracy	Sensitivity/Recall	Precision	Specificity	F1-Score
M1_Pool 1,2	86.45%	86.45%	88.21%	39.89%	87.29%
M1_Pool 2,3	86.47%	86.47%	88.23%	39.98%	87.31%
M1_Pool 3,4	86.48%	86.48%	88.24%	39.99%	87.32%
M1_Pool 4,5	86.49%	86.49%	88.26%	40.08%	87.34%
M2_Pool 1,2,3	86.51%	86.51%	88.27%	40.11%	87.35%
M2_Pool 2,3,4	86.66%	86.66%	88.41%	40.26%	87.50%
M2_Pool 3,4,5	87.52%	87.52%	88.92%	42.53%	88.13%

Similarly, among the M2 configurations, a clear pattern emerges: models with deeper pooling layers consistently outperform those with shallower models. With accuracy of 87.52%, sensitivity/recall of 87.52%, precision of 88.92%, specificity of 42.53%, and an F1-score of 88.13%, M2_Pool 3,4,5, outperforms across all metrics. This emphasises how deeper feature representations can capture more complex patterns in the data, leading to better classification robustness and accuracy.

Table 9.7 Performance metrics evaluated for MHMA System on Subset 3, integrating the Attention Model with Mod_AlexNet utilizing the RMSProp optimizer at a learning rate of 0.0001, 50 epochs, and batch size of 64.

Model/FC Layer	Accuracy	Sensitivity/Recall	Precision	Specificity	F1-Score
M1_Pool 1,2	85.64%	85.64%	87.23%	37.22%	86.41%
M1_Pool 2,3	85.75%	85.75%	87.37%	37.83%	86.53%
M1_Pool 3,4	86.00%	86.00%	87.39%	37.83%	86.67%
M1_Pool 4,5	87.13%	87.13%	87.54%	37.96%	87.31%
M2_Pool 1,2,3	86.12%	86.12%	87.47%	38.02%	86.76%
M2_Pool 2,3,4	86.18%	86.18%	87.57%	38.56%	86.83%
M2_Pool 3,4,5	86.25%	86.25%	87.72%	39.47%	86.94%

Table 9.7 presents the performance metrics for MHMA System on Subset 3, incorporating the Attention Model and Mod_AlexNet architecture with RMSProp optimiser. Across various configurations denoted by M1_Pool and M2_Pool, the classification capabilities of the model are evaluated. Performance metrics show an improvement with deeper pooling layers, starting with the M1 configurations. The model obtains 85.64% accuracy, 85.64% sensitivity/recall, 87.23% precision, 37.22%

specificity, and 86.41% F1-score for M1_Pool 1,2. Performance steadily improves with increasing pooling depth and peaks in M1_Pool 4,5 with F1-score values of 87.31%, 87.13%, 87.54%, 37.96%, and accuracy, sensitivity/recall, precision, and specificity, respectively. This implies that improved classification accuracy and robustness are a result of deeper feature representations.

Observing the M2 configurations, performance measures improve with increasing pooling layer depth. Among all M2 configurations, M2_Pool 3,4,5 notably performs most effectively, with F1-score values of 86.94%, 87.72%, 39.47%, and 86.25% for accuracy, sensitivity/recall, precision, and specificity, respectively.

Analysing the data in Tables 9.5, 9.6, and 9.7 offers new perspectives on how MHMA System functions on Subset 3 with various optimiser setups. On a batch size of 64, each table shows the integration of the Mod_AlexNet and Attention Model with the various optimisers, SGDM, Adam, and RMSProp, respectively. All three optimiser setups are often surpassed by M2_Pool 3,4,5, which achieves the optimum performance of all the configurations examined in each table. Out of all the optimisers, the SGDM optimiser performed the best. M2_Pool 3,4,5 produced the best results, with an accuracy of 90.99%, specificity of 51.99%, and F1-score of 91.26%. This configuration shows the best results across all performance measures when compared to other configurations and different optimisers.

9.3.3 Batch Size 128

Tables 9.8, 9.9, and 9.10 explore and analyse the results acquired by training MHMA System using a learning rate of 0.0001 across 50 epochs and a batch size of 128, employing SGDM, Adam, and RMSProp optimisers, respectively.

The performance metrics of MHMA System on Subset 3, which uses the SGDM optimiser with a learning rate of 0.0001, 50 epochs, and a batch size of 128 and incorporates the Attention Model with Mod_AlexNet, are shown in Table 9.8. Consistent improvements can be found in accuracy, sensitivity/recall, precision, specificity, and F1-score measures across different pooling layer configurations. The model achieves an accuracy of 84.68%, sensitivity/recall of 84.68%, precision of 87.38%, specificity of 36.91%, and F1-score of 85.95%, starting with M1_Pool 1,2. Performance significantly improves as the model moves through the pooling layers,

with M1_Pool 4,5 reporting the greatest results (85.07% accuracy, 85.07% sensitivity/recall, 87.67% precision, 37.36% specificity, and 86.29% F1-score).

Table 9.8 Performance metrics evaluated for MHMA System on Subset 3, integrating the Attention Model with Mod_AlexNet utilizing the SGDM optimizer at a learning rate of 0.0001, 50 epochs, and batch size of 128.

Model/FC Layer	Accuracy	Sensitivity/Recall	Precision	Specificity	F1-Score
M1_Pool 1,2	84.68%	84.68%	87.38%	36.91%	85.95%
M1_Pool 2,3	85.01%	85.01%	87.60%	36.91%	86.22%
M1_Pool 3,4	85.02%	85.02%	87.61%	37.06%	86.24%
M1_Pool 4,5	85.07%	85.07%	87.67%	37.36%	86.29%
M2_Pool 1,2,3	84.81%	84.81%	87.64%	37.05%	86.14%
M2_Pool 2,3,4	85.26%	85.26%	87.78%	37.37%	86.44%
M2_Pool 3,4,5	85.53%	85.53%	88.01%	37.39%	86.68%

As the model moves to the M2 models, there are further gains in performance in every metric when compared to the M1 models. M2_Pool 1,2,3, for example, exhibits a small improvement over the related M1 model with accuracy of 84.81%, sensitivity/recall of 84.81%, precision of 87.64%, specificity of 37.05%, and F1-score of 86.14%. A similar 85.53% accuracy, 85.53% sensitivity/recall, 88.01% precision, 37.39% specificity, and 86.68% F1-score are recorded by M2_Pool 3,4,5.

Table 9.9 Performance metrics evaluated for MHMA System on Subset 3, integrating the Attention Model with Mod_AlexNet utilizing the Adam optimizer at a learning rate of 0.0001, 50 epochs, and batch size of 128.

Model/FC Layer	Accuracy	Sensitivity/Recall	Precision	Specificity	F1-Score
M1_Pool 1,2	83.07%	83.07%	87.31%	32.83%	85.10%
M1_Pool 2,3	83.15%	83.15%	87.33%	33.95%	85.15%
M1_Pool 3,4	83.17%	83.17%	87.35%	34.15%	85.17%
M1_Pool 4,5	82.94%	82.94%	87.37%	34.57%	85.05%
M2_Pool 1,2,3	82.92%	82.92%	87.33%	34.56%	85.03%
M2_Pool 2,3,4	83.32%	83.32%	87.41%	35.03%	85.27%
M2_Pool 3,4,5	83.80%	83.80%	87.46%	36.25%	85.56%

Table 9.9 presents an evaluation of different models in MHMA System, which combines Mod_AlexNet with the Attention Model, and provides noteworthy information about their performance measures. All models were evaluated on Subset

3 utilising the Adam optimiser, with a learning rate of 0.0001, 50 epochs, and a batch size of 128. M2_Pool 3,4,5 leads in overall performance with an accuracy of 83.80% and a close second in precision at 87.46%. M1_Pool configurations show a compromise between accuracy (83.07% - 83.17%) and precision (87.31% - 87.37%). This dynamic is highlighted by the fact that M1_Pool 1,2 has the lowest accuracy but the highest precision. With respect to TPs and TNs, M2_Pool 3,4,5 continues to be strong with respect to specificity (36.25%), but oddly, it has the lowest sensitivity/recall (83.80%). M1_Pool models, on the other hand, exhibit reduced specificity but stable sensitivity/recall. They have trouble correctly classifying negatives, but they are excellent at spotting actual positives.

The proportion of correctly identified instances, or the accuracy, shows a progressive increase across the configurations. The accuracy of M2_Pool 1,2,3 is 82.92%; M2_Pool 2,3,4 is 83.32%; and M2_Pool 3,4,5 is 83.80%. This pattern implies that deeper pooling layer configurations could result in predictions from the model that are more accurate overall. Consistent developments are revealed by further investigation of precision, F1-score, and sensitivity/recall for all pooling layer configurations. Sensitivity/recall, which ranges from 82.92% to 83.80%, is comparatively constant across configurations. Additionally, precision, which is defined as the percentage of accurately detected positive cases among all cases classified as positive, varies barely, ranging between 87.33% and 87.46%. This suggests that the models continue to accurately detect positive instances irrespective of the architecture of the pooling layer. Furthermore, there is a minor increase in specificity with deeper pooling layers, indicating the ability of the model to accurately identify negative cases. The highest specificity is achieved by M2_Pool 3,4,5, at 36.25%, while M2_Pool 1,2,3 reaches 34.56%. This means that the ability of the model to distinguish between instances that are positive and negative may be improved by adding more pooling layers.

The performance indicators for MHMA System on Subset 3 using the Mod_AlexNet architecture and RMSProp optimiser with the Attention Model are shown in Table 9.10. Seven configurations, identified as M1_Pool 1,2, M1_Pool 2,3, M1_Pool 3,4, M1_Pool 4,5, M2_Pool 1, 2, 3, M2_Pool 2, 3, 4, and M2_Pool 3, 4, 5 are included in the table with different pooling layers.

In Table 9.10, for M1 model, across all configurations, accuracy varies from 84.08% to 84.20%. This suggests that throughout the various pooling layer thicknesses, overall classification accuracy performance is rather stable. The accuracy values are notable for being somewhat lower than those achieved with other optimisers, such as Adam or SGDM, indicating that RMSProp may not be as useful in this specific instance. Sensitivity/recall, with values ranging between 84.08% and 84.20%. Precision also shows little fluctuation, with values between 86.02% and 86.08%. With scores ranging from 24.27% to 24.64%, specificity, a measure of the accuracy of the model in identifying negative instances, remains comparatively low across all configurations. This is a possible area for development since it implies that the models trained with RMSProp have more difficulty correctly detecting negative cases than positive instances.

Table 9.10 *Performance metrics evaluated for MHMA System on Subset 3, integrating the Attention Model with Mod_AlexNet utilizing the RMSProp optimizer at a learning rate of 0.0001, 50 epochs, and batch size of 128.*

Model/FC Layer	Accuracy	Sensitivity/Recall	Precision	Specificity	F1-Score
M1_Pool 1,2	84.08%	84.08%	86.02%	24.27%	85.03%
M1_Pool 2,3	84.08%	84.08%	86.02%	24.27%	85.03%
M1_Pool 3,4	84.12%	84.12%	86.04%	24.45%	85.07%
M1_Pool 4,5	84.20%	84.20%	86.08%	24.64%	85.12%
M2_Pool 1,2,3	84.15%	84.15%	86.05%	24.46%	85.09%
M2_Pool 2,3,4	84.21%	84.21%	86.10%	24.70%	85.14%
M2_Pool 3,4,5	84.26%	84.26%	86.16%	25.12%	85.20%

The accuracy range in Tale 9.10 for the M2 model is 84.15% to 84.26% in all configurations, suggesting a consistent performance in terms of overall classification accuracy. This implies that the ability of the model to accurately classify examples is mostly unaffected by the combinations of pooling layers. It is important to note that these accuracy numbers are marginally lower than those of other optimiser settings, which may indicate that RMSProp is less successful in this scenario in terms of attaining high overall accuracy. Sensitivity/recall, with values between 84.15% and 84.26%. This suggests that irrespective of the pooling layer architecture employed, the ability of the model to identify positive cases stays consistent. There is little variance in precision across configurations, with values ranging from 86.05% to

86.16%, indicating the accuracy of the model in identifying positive cases out of all instances classified as positive. This implies that the accuracy of the model in classifying positive occurrences is not greatly affected by the combinations of pooling layers. However, specificity, which measures the ability of the model to properly detect negative cases, increases slightly as the number of pooling layers grows. In particular, specificity varies from 24.46% to 25.12%, suggesting that models with combinations of deeper pooling layers typically outperform models with less pooling layers in terms of accurately identifying negative cases.

Analysing the data from Tables 9.8, 9.9, and 9.10 reveals new insights into the performance of the MHMA System on Subset 3 under various optimiser setups. With a batch size of 128 in each table, the attention model and Mod_AlexNet are integrated with various optimisers, namely SGDM, Adam, and RMSProp. Notably, the M2_Pool 3,4,5 arrangement consistently performs better than other setups, as evidenced by superior performance metrics across all tables. In particular, M2_Pool 3,4,5 has the highest accuracy (85.53%), specificity (37.39%), and F1-score (86.68%) among all configurations and optimiser combinations, indicating that SGDM is the most successful optimiser out of the three.

9.3.4 Optimal Outcomes Analysis

Figures 9.4 and 9.5 demonstrate training and validation curves that represent the accuracy and loss of attention model M2_Pool3,4,5 with batch size of 64, and over 50 epochs. Figure 9.4 shows the evolution of training and validation accuracy through epochs. Training accuracy rapidly increased from 45% to 99%, eventually stabilising at the peak. Additionally, the training accuracy becomes nearly constant after the 20^{th} epoch. Validation accuracy, on the other hand, starts at 65% and continuously increases until it reaches 93%. Following the 25^{th} epoch, the validation accuracy approaches a near-constant state. In Figure 9.5, which displays the training and validation losses, the training loss starts at 2.46 and steadily decreased, reaching a stable 0.004 by the 20^{th} epoch and remaining particularly constant until the 50^{th} epoch. On the other hand, the validation loss starts at 2.16 and gradually decreases, reaching a minimal loss of 0.01 by the 20^{th} epoch and remaining rather steady until the 50^{th} epoch.

A more thorough analysis of the data in Tables 9.2 through 9.10 confirms the effectiveness of a batch size of 64 and supports the superiority of SGDM over alternative optimisers. Surprisingly, using the SGDM optimiser with a 64-batch batch size produces the most optimum results, with 90.99% accuracy, 91.98% precision, and 91.26% F1-score. This model stands out for having a significant increase in specificity, indicating that it is more capable of identifying negative cases than other models. The results highlight how important optimiser choice and batch size are to improving the performance of the MHMA System on Subset 3. In addition to demonstrating the advantages of the SGDM optimiser, the results highlight the need of using an ideal batch size in order to achieve better model performance, especially with regard to accuracy, precision, and specificity.

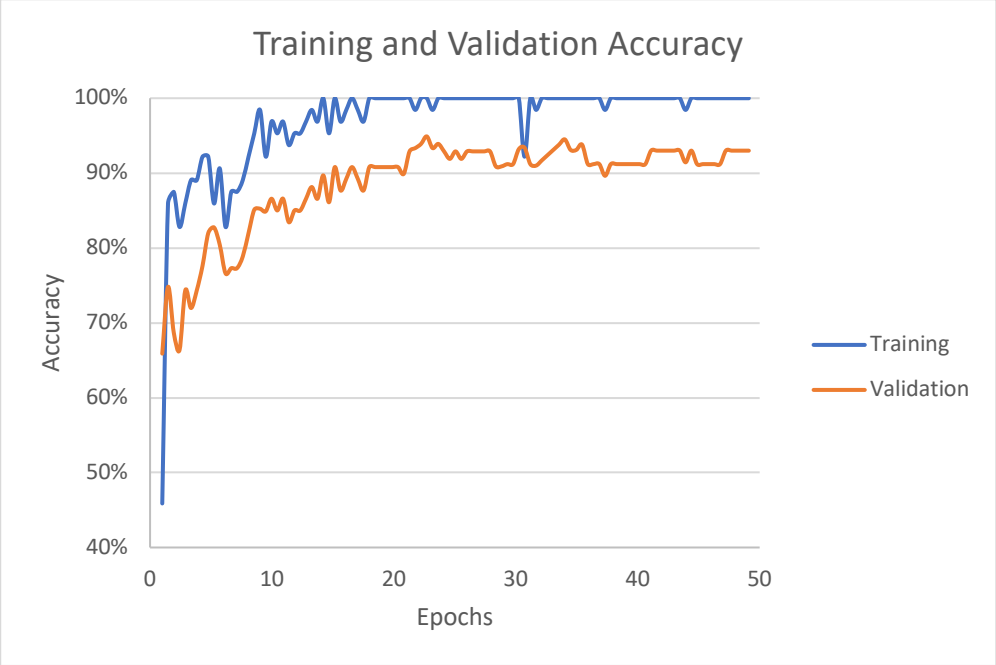

Figure 9.4 Self-Attention Training and Validation Accuracy vs Epochs for M2_Pool 3,4,5

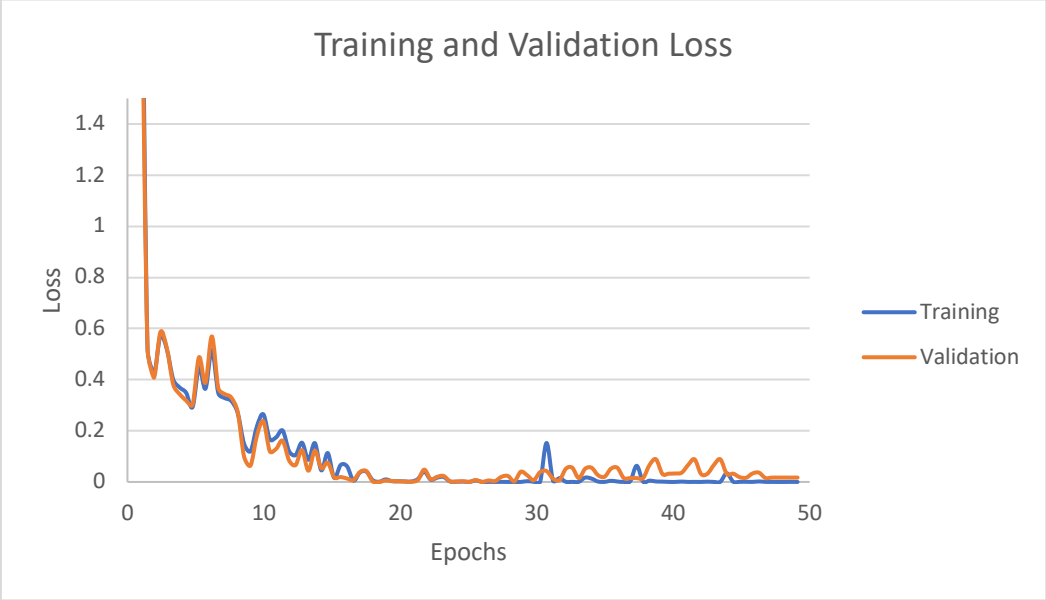

Figure 9.5 Self-Attention Training and Validation Loss vs Epochs for M2_Pool 3,4,5

Cohen's d measure and t-test, two statistical significance tests, were used to assess the efficacy of the MHMA System, as mentioned in Chapter 5, and measure the impact of the third contribution, the addition of the attention model. These measurements were used to evaluate the rates of improvement between the optimal performance of the MHMA System and the performance of the MA System. The main goal of calculating the improvement rates is to determine how the third contribution, integrating the developed attention model, affects deep learning performance. Notably, the modified Mod_AlexNet was employed as the deep learning model in this system.

A comprehensive analysis of performance metrics and the measure of improvement between the reference MA System performance and MHMA System implementations with varying batch sizes are shown in Table 9.11. The F1-score, accuracy, sensitivity, precision, and specificity are among the performance metrics taken into consideration.

In contrast to MA System, MHMA System performs not as well when analysing data with a batch size of 32 in terms of accuracy, sensitivity, and specificity. On the other

hand, accuracy and the F1-score indicate a minor improvement. A small effect size is indicated by Cohen's d value for accuracy, pointing to a moderate difference between the two systems. The T-test significance result, however, indicates that there may not be a statistically significant improvement. In contrast, MHMA System with a 64-batch size exhibits improved F1-score, accuracy, sensitivity, precision, and specificity when compared to MA System. A medium impact size is indicated by Cohen's d value for accuracy, pointing to a more significant difference between the two systems. Additionally, MHMA System exhibits inconsistent performance in comparison to MA System with a batch size of 128. Precision and the F1-score show slight improvements, while accuracy, sensitivity, and specificity all decline. A small effect size is indicated by Cohen's d value for accuracy, pointing to a marginal difference between the two systems. Once more, this difference is not statistically significant according to the T-test significance value.

A detailed comparison of the SGDM, Adam, and RMSProp optimizers was conducted across batch sizes of 32, 64, and 128. Results indicated that SGDM consistently outperformed Adam and RMSProp in terms of accuracy, F1-score, and specificity. This can be attributed to SGDM's ability to better manage learning rate adjustments, which reduces the risk of overfitting and allows more stable convergence.

In conclusion, the configuration with a batch size of 64 stands out as the most promising among the various batch sizes examined for MHMA System. It continuously outperforms MA System in every metric, with significant gains in F1-score, accuracy, sensitivity, and precision. The major improvement in specificity, which shows a significant improvement in correctly identifying negative situations, is especially remarkable. This statistically significant increase in specificity points to a considerable improvement in performance. Based on these findings, MHMA System with a batch size of 64 therefore stands out as a solid candidate for more thought and execution, demonstrating its capacity to successfully handle the challenges raised by MA System in the classification of abnormal cases. MA System exhibits a specificity value of 23.91%, indicating its capability to accurately identify negative cases. In contrast, MHMA System, employing a batch size of 64, demonstrates a substantial improvement in specificity, reaching a value of 51.99%. This notable

increase underscores the enhanced ability of MHMA System to effectively distinguish negative cases, reflecting its superior performance compared to MA System with an improvement rate of 117% in the specificity measure.

Table 9.11 Comparative Evaluation of Performance Metrics and Significance of Improvement Using Cohen's d and T-test Significance for MHMA System Implementation with Different Batch Sizes in Comparison to MA System Performance

System	Performance Measures					Cohen's d:		T-test (p)	
	Accuracy	Sensitivity	Precision	Specificity	F1-Score	Measure	Significance	Measure	Significance
MA System	91.61%	91.61%	88.16%	23.91%	89.57%	-	-	-	-
MHMA System with a Batch Size of 32	86.46%	86.46%	88.21%	39.89%	87.30%	0.0269	Small	0.8693	Not Significant
MHMA System with a Batch Size of 64	90.99%	90.99%	91.98%	51.99%	91.26%	0.2651	Medium	0.30203	Not Significant
MHMA System with a Batch Size of 128	85.53%	85.53%	88.01%	37.39%	86.68%	-0.013172	Small	0.92904	Not Significant

While the MHMA model shows strong sensitivity and accuracy, specificity remains a challenge, indicating a higher rate of false positives. In clinical practice, false positives can result in unnecessary biopsies and patient stress. To improve specificity, future work will explore threshold tuning and ensemble strategies to refine decision boundaries and reduce misclassifications.

9.4 Summary

In this chapter, the use of the developed deep learning model Mod_AlexNet in conjunction with an entirely novel Multi-Head Attention model was introduced and examined. This combined method addresses the three primary challenges that arise

in the detection of breast cancer, namely: poor performance in multi-class classification situations, the complexity of different breast sizes and densities, and the crucial role of differentiating between benign and malignant abnormalities. Work in this chapter heavily depends on the newly developed Multi-Head Attention Model, that operates in cooperation with the previous Mod_AlexNet to significantly improve its performance, especially in correctly classifying abnormal classes. The Multi-Head Attention model has a number of significant advantages. One of these is its ability to focus on relevant areas in tomosynthesis images, which helps to reduce the complexity caused by differences in breast anatomy. This model efficiently obtains discriminative information necessary for accurate classification by giving selective attention to key traits, making it easier to distinguish benign from malignant abnormalities. The diagnostic efficacy of Mod_AlexNet in breast cancer detection tasks is further enhanced by the ability of the Multi-Head attention model to capture complex spatial dependencies and correlations among image regions, which allows for a deeper informative feature extraction.

The effectiveness of the MHMA model is carefully examined and compared to the performance of Mod_AlexNet in the results and discussion section. A wide range of performance indicators are carefully evaluated, such as the F1-Score, accuracy, sensitivity, precision, and specificity. The results show significant improvements in the performance of the MHMA System compared to Mod_AlexNet, especially when the batch size is 64. The statistical significance of these observed benefits is determined using statistical analysis utilising techniques like Cohen's d and T-tests, highlighting the ways in which the incorporation of the Multi-Head attention model enhances the effectiveness of deep learning models. Additionally, the FFS-EC, introduced in Chapter 8, highlights the importance and impact of ensemble modelling, feature fusion, and selection, which leads to the development of the final system. The enhancements from previous systems are incorporated and a new contribution to ensemble modelling is developed. The next chapter will explore the details of the final system, explaining its design and examining its findings.

Chapter 10 Hybrid Multi-Head Attention-Based System with Feature Fusion and Ensemble Classification Model

10.1 Introduction

The Hybrid Multi-Head Self-Attention Model with Feature Fusion, Selection, and IVECM for Enhanced DBT Classification System, HMSA-IVECM System concludes with the introduction of an innovative ensemble model. While earlier systems, MA, FFS-EC, and MHMA systems, each made a significant contribution in addressing challenges related to multi-class classification accuracy, varying breast density and size and abnormal classification, limited performance in the abnormal classes classification still persists.

Through the smooth integration of features from FFS-EC System and MHMA System, HMSA-IVECM System overcomes the challenges presented by differences in breast densities and sizes. This increase is specifically remarkable for multi-class classification, where it shows significant improvements, especially in abnormal classes. Interestingly, substantial enhancements in the classification performance for both benign and malignant classes have been observed, outperforming the performance of traditional deep learning models used for the classification of tomosynthesis scans.

10.2 Methodology

The HMSA-IVECM system starts with image augmentation, followed by image enhancement and colour mapping as outlined in previous chapters, with a comprehensive explanation provided in Chapter 3. For feature extraction, a combination of Mod_AlexNet and the Multi-Head attention model is employed alongside HOG descriptors. By concatenating features with HOG descriptors, this integrated model acquires and fuses features, enhancing the model with relevant and complex features that are recognised for their robustness. This improves inter-class discrimination by enriching class-specific information. Following feature fusion, three powerful feature selection methods (mRMR, chi-square, and f-test) are used to identify significant features, utilising the innovations presented by FFS-EC.

The selected features are then classified using five high-performing classifiers: SVM, NB, DT, FDA, and KNN, inspired by the significant improvements noticed in the results of the FFS-EC System. In addition, an ensemble model was employed to improve classification performance. Two predictions are generated by the ensemble model: classifier weights representing the F1-score weighted average and the specificity weighted average. The final multi-class prediction is obtained by combining these predictions—which come from applying three feature selection techniques—with a maximum voting ensemble model. The architecture of HMSA-IVECM System is illustrated in Figure 10.1.

Furthermore, every model in the tomosynthesis classification has distinctive advantages that add to its effectiveness. Robust feature extraction capabilities are provided by the Mod_AlexNet combined with the Multi-Head attention model, which captures complex patterns essential for precise classification. One of the key advantages of the self-attention mechanism within the HMSA-IVECM System is its ability to improve model interpretability. By generating attention heatmaps, the model can visually highlight areas within the tomosynthesis scans that are most influential in classification. In addition, HOG descriptors improve feature richness by capturing texture and shape information. The ensemble method combines different classifiers and takes advantage of the unique capabilities of each classifier to reduce weaknesses and improve classification performance. Model efficiency and interpretability are improved by feature selection approaches like MRMR, chi-square, and f test, which guarantee that only the most discriminative features contribute to the classification process. The use of multiple feature selection methods (mRMR, chi-square, and f-test) was designed to capture complementary information from the fused feature sets. This improves the richness of extracted features and reduces complexity of the feature set.

Additionally, the ensemble model promotes cooperation amongst several predictions, utilising collective intelligence to improve the final classification decision and boosting the ability of the system to adapt and reliability.

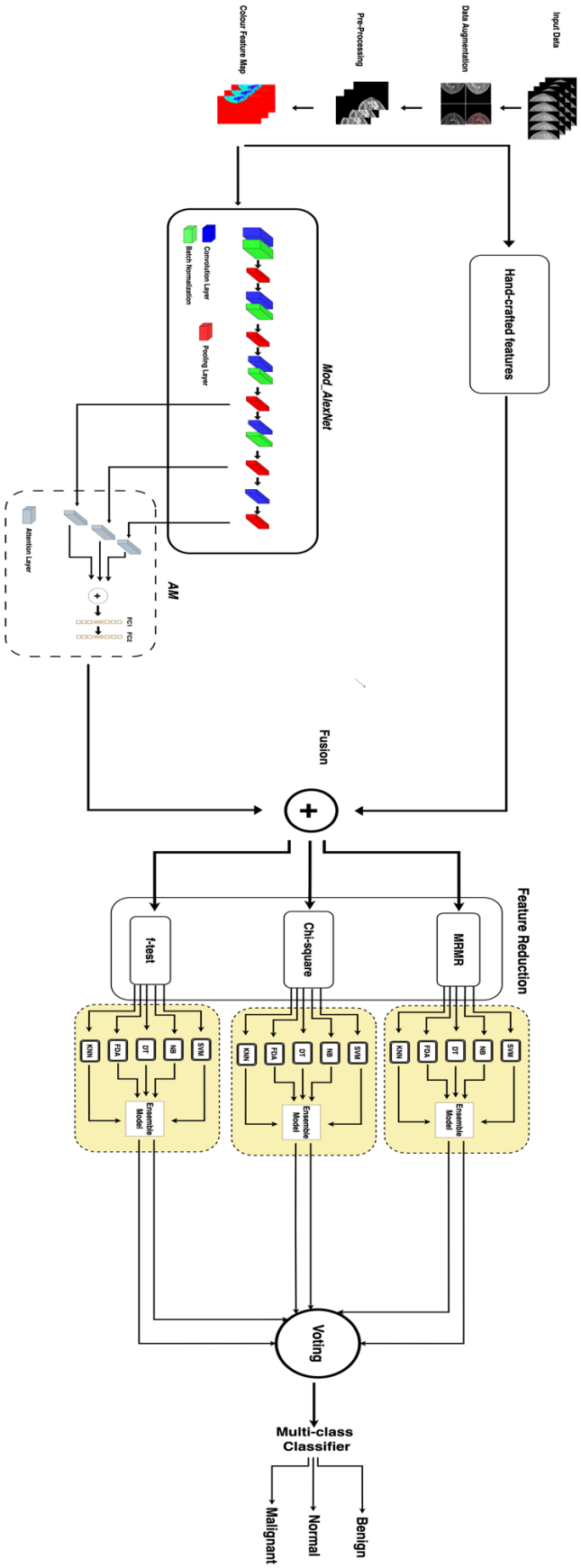

Figure 10.1 Architecture of HMSA-IVECM System

An ensemble model specifically designed to improve the classification of abnormal classes—a critical component in the analysis of tomosynthesis scans—is presented in this work as a novel contribution. Two essential elements are introduced in this attempt by the research: class weights and classifier weights. These weights are used to indicate, respectively, the relative relevance of various classes in the classification task and the weights given to classifiers in the ensemble. This contribution utilises the stacking ensemble approach and analyses its methodology, which combines the predictions of several base classifiers. The goal of this technique is to overcome the weaknesses of individual classifiers and improve overall predicted accuracy and robustness by combining the outputs of different classifiers via a meta-classifier.

Using a hierarchical architecture, the stacking ensemble model consists of several layers of classifiers cooperating to generate predictions. Using the given training dataset, a varied ensemble of base classifiers is trained to begin the process. Each base classifier is trained on the same subset of the data. Due to the complex nature of tomosynthesis data, where different abnormalities exhibit distinctive properties that are well recorded by diverse classifiers, this diversity is especially important for tomosynthesis scan classification. The trained base classifiers then produce predictions for all cases in the validation set on their own. These predictions serve as input features for the meta-classifier, which is tasked with integrating them optimally to produce the final prediction

In this chapter, class and classifier weights are developed and applied to provide a new contribution over the traditional ensemble model. By adding more precise weight assignments for classifiers and classes, this development attempts to increase the efficacy of classification. Specifically, the introduction of class weights addresses the common issue of class imbalance in the dataset under consideration. Given the unbalanced distribution of normal, benign, and malignant cases, adjusted class weights help to correct the imbalance by giving more weight to underrepresented classes. By making this strategic change, the discriminative abilities of the ensemble classifier are strengthened, and every class is guaranteed to have an equal impact on its decision-making process.

The corresponding lack or abundance of images within each class is carefully considered while determining class weights. Classes that have a few instances are given larger weights to make up for their underrepresentation, which guarantees that they contribute fairly to the decision-making of the ensemble classifier. By enabling the ensemble classifier to adjust to the different class dominance in the dataset, this adaptive weighting technique reduces the negative effects of class imbalance and promotes more fair and clinically meaningful predictions. A critical step in resolving class imbalance in multi-class datasets is the determination of class weights using a prescribed formula:

$$w_i = \frac{n_s}{n_c * n_{si}} \quad (10.1)$$

Where w_i denotes the weight assigned to each class, with i representing the specific class. The numerator n_s denotes the total number of samples in the dataset, providing a measure of the overall size of the dataset. Conversely, n_c indicates the total number of unique classes within the target variable, illustrating the diversity of classes present in the dataset. Finally, n_{si} signifies the total number of instances associated with the respective class i. Based on class frequency in relation to dataset size, this formula works by inversely changing weights. The model achieves increased sensitivity to the unique features of every class by assigning higher weights to classes with fewer instances and lower weights to those with greater representation. This promotes a more balanced learning process.

Following the assignment of class weights to each prediction, these predictions are fed into a meta classifier, representing a substantial advancement in multi-class classification approaches. This novel method creates a complex mechanism for producing weighted results by combining classifier predictions with performance measurements like specificity and the F1 score. The F1 score provides a comprehensive assessment of classifier performance as a composite of precision and recall, while specificity measures the ability to correctly detect TNs. Through the integration of these performance indicators into the weighting scheme, the meta classifier enhances the predictive accuracy by giving more weight to predictions made by classifiers with higher F1 scores and specificity.

This innovative contribution has significant implications for the field of multi-class classification challenges. Through the combined use of performance indicators and classifier predictions, the meta classifier presents a new framework for decision-making. It enables the combining of weighted predictions from several classifiers for each class, therefore utilising the collective intelligence embedded in various prediction techniques. Moreover, this methodology has remarkable flexibility and can be easily integrated with a wide range of classifiers, which allows the development of ensemble models that take use of the various benefits of individual classifiers while also efficiently addressing their limitations.

The final phase of this methodology involves implementing a majority voting ensemble model into use, which is a critical step in improving prediction accuracy. This method involves adding together the predictions generated by every base classifier and feature selection model and then selecting the class label that receives the most votes as the final prediction. This ensemble model follows an equal decision-making procedure in which every prediction is given the same weight in deciding the final classification result. When the ensemble model receives predictions, it methodically accumulates the votes for each class label across all predictions before selecting the label with the highest vote count as the final prediction. Figure 10.2 provides an architectural representation of this newly developed Integrated Voting Ensemble Classification Model, IVECM, which is an important contribution.

Figure 10.2 demonstrates the integration process within the classification system. Three distinct feature selection models generate classifier outputs that are fed into a class weight model, which assigns a weight to each prediction based on its class. These weighted predictions are then fed into a meta classifier, which gives each classifier output two weights. These weights indicate the f1-score and specificity, respectively. Furthermore, each feature selection model provides two predictions: one based on the f1-score for each classifier, and another based on the specificity score. To get to the final prediction, the output of the meta classifier is fed into a maximum voting model.

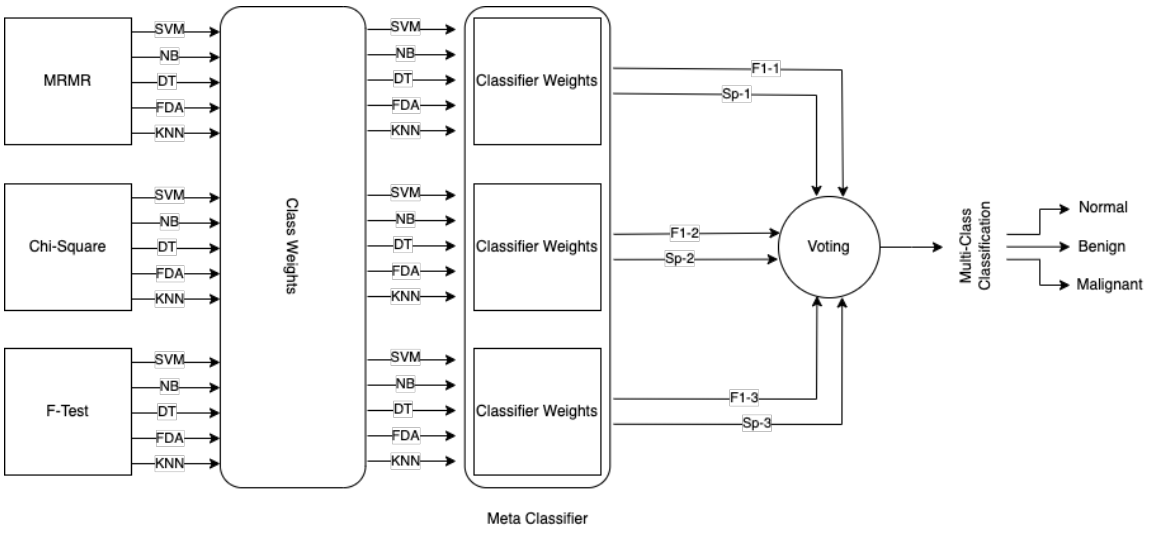

Figure 10.2 Architecture of the IVECM

The proposed meta classifier has many benefits over traditional stacked ensemble models, which mostly use methods like voting or averaging to combine base classifier outputs. Most importantly, it makes sure that predictions are carefully weighted based on classifier confidence and performance by integrating performance measures into the weighting system. This leads to better informed decision-making processes. Moreover, having the ability to combine many classifiers and performance indicators allows for more precise control over the weighting procedure, which results in predictions that are noticeably more accurate and consistent. To sum up, the development of this meta classifier is a major advancement, providing an improved prediction accuracy. To simplify the ensemble model for clinical implementation, a weighted average voting system was employed, where classifiers with higher specificity and F1 scores were given greater influence in the final prediction. This ensures that the ensemble output reflects the strengths of individual classifiers while minimizing inconsistencies. The model can also be pruned to reduce complexity by limiting the number of base classifiers without sacrificing predictive accuracy. A simplified version of the ensemble could improve adoption in clinical settings while maintaining high diagnostic performance.

The following section outlines the implementation, the results that followed, and their discussion regarding HMSA-IVECM System. This section provides a thorough explanation of how the system was deployed, describes the outcomes that were achieved, and dives deeply into the evaluation and interpretation of these results.

10.3 Results and Discussion

In this study, the HMSA-IVECM System is evaluated, with a focus on the concluding contribution, the ensemble model. This section includes a thorough examination of the results obtained from the deployment of HMSA-IVECM System. A comprehensive evaluation was carried out to determine the efficacy of the integration of the ensemble model with Mod_AlexNet, the developed Attention Model, feature fusion, and feature selection methods. Performance metrics like the F1 score, sensitivity, specificity, precision, and accuracy were all included in the analysis. The performance of the HMSA-IVECM System was compared to that of MA, FFS-EC and MHMA Systems. A t-test and Cohen's d were used to determine the statistical significance of any observed improvements.

The initial stage of this system consisted of integrating the HOG descriptors with features extracted from the developed Attention Model, which corresponds to MHMA System. Concatenation was used as the integration method in this fusion procedure.

Table 10.1 Performance Assessment of HMSA-IVECM System fusing Features from SelfAttention_ModAlexNet and HOG Descriptors Across Multiple Classifiers

Different integrated contexts	Classifier	Performance Measure				
		Accuracy	Sensitivity/Recall	Precision	Specificity	F1-Score
SelfAttention_ModAlexNet training on SGDM		90.99%	90.99%	91.98%	51.99%	91.26%
Feature Fusion with HOG descriptors	**NB**	90.89%	90.89%	93.55%	53.55%	91.40%
	SVM	90.35%	90.35%	93.64%	55.04%	91.12%
	DT	89.66%	89.66%	92.19%	54.39%	90.58%
	FDA	92.69%	92.69%	92.59%	47.87%	92.41%
	KNN	90.80%	90.80%	93.40%	54.05%	91.40%

The performance evaluation of HMSA-IVECM System is shown in Table 10.1, where several classifiers were employed to fuse features extracted from the SelfAttention_ModAlexNet with HOG descriptors. With corresponding classifiers and performance metrics such as accuracy, sensitivity/recall, precision, specificity, and F1-Score, each row represents a unique integrated context.

When comparing the performance of MHMA System (SelfAttention_ModAlexNet) with that of the fusion system that combines features extracted from the MHMA

System with HOG descriptors for several classifiers, as can be seen by examining Table 10.1, a significant improvement in the specificity measure is seen for all classifiers except for the FDA classifier. The FDA classifier, in particular, reported the lowest specificity rate of 47.87% and the highest accuracy rate of 92.69%, demonstrating that it was more effective in classifying normal cases than abnormal ones. Further investigation demonstrated that the performance of the NB, SVM, and KNN classifiers followed similar patterns, with minor differences. Remarkably, the SVM classifier achieved the greatest precision rate of 93.64% and the highest specificity rate of 55.04%, outperforming the NB classifier. On the other hand, the F1-Score that the KNN and NB classifiers achieved was 91.40%, which is the second highest of all the classifiers. The DT classifier additionally demonstrated the lowest accuracy rate of 89.66%, the second-highest specificity rate of 54.39%, the precision rate of 92.19%, and an F1-Score of 90.58%. Overall, the results highlight the accuracy of the employed classifiers. A comparison with the performance of the MHMA System indicates that the fusion model greatly improves measures of specificity and precision, allowing for more accurate classification of abnormal cases.

In order to reduce the feature set and enhance performance, three feature selection techniques—MRMR, chi-square test, and f-test—were assessed after the features were fused. Tables 10.2, 10.3, and 10.4 in this chapter present the results of the comparative analysis of several classifiers.

Table 10.2 Performance Evaluation of HMSA-IVECM System Employing the MRMR Feature Selection Model with Various Classifiers

Different integrated contexts	Classifier	Performance Measure				
		Accuracy	Sensitivity/Recall	Precision	Specificity	F1-Score
SelfAttention_ModAlexNet training on SGDM		90.99%	90.99%	91.98%	51.99%	91.26%
Feature Selection (MRMR)	**NB**	89.16%	89.16%	91.03%	50.36%	89.94%
	SVM	90.41%	90.41%	93.73%	55.07%	91.16%
	DT	89.23%	89.23%	92.31%	57.15%	90.37%
	FDA	92.99%	92.99%	92.82%	45.17%	92.54%
	KNN	93.10%	93.10%	93.81%	53.88%	92.97%

After evaluating the performance metrics for each classifier in Table 10.2, it appears that the SVM and KNN classifiers often outperform the others across a variety of measures. With a precision of 93.73%, accuracy of 90.41%, sensitivity of 90.41%, specificity of 55.07%, and F1-score of 91.16%, the SVM classifier performs optimally. Comparably, the KNN classifier performs well on all metrics, with the best accuracy of 93.10% and a remarkable F1-score of 92.97%. On the other hand, the F1-score values and accuracy of the NB and DT classifiers are relatively lower, indicating possible limitations in their classification skills. The DT classifier performs worse than SVM and KNN in terms of accuracy (89.23%) and precision (92.31%), although having a comparatively high specificity (57.15%). On the other hand, its F1-score and sensitivity/recall values are similar to those of SVM, suggesting that it can accurately recognise both positive and negative examples. In contrast, the accuracy (89.16%) and precision (91.03%) of the NB classifier are lower than those of SVM and KNN. On the other hand, NB has comparable sensitivity/recall values to SVM, suggesting that it can recognise genuine positive cases. In contrast to SVM and KNN, NB has a lower specificity (50.36%), which raises the possibility of FPs. The FDA classifier, only records 45.17% in terms of specificity, suggesting a larger number of FPs, even though it achieves the highest total accuracy of 92.99%.

The results presented in Table 10.3, which outline the performance evaluation of HMSA-IVECM System after implementing the Chi-square test feature selection model with various classifiers, show that the performance of each classifier varies slightly across different performance measures.

Table 10.3 Performance Evaluation of HMSA-IVECM System Employing the Chi-Square test Feature Selection Model with Various Classifiers

Different integrated contexts	Classifier	Performance Measure				
		Accuracy	Sensitivity/Recall	Precision	Specificity	F1-Score
SelfAttention_ModAlexNet training on SGDM		90.99%	90.99%	91.98%	51.99%	91.26%
Feature Selection (Chi-Square test)	**NB**	91.12%	91.12%	93.49%	53.19%	91.57%
	SVM	90.44%	90.44%	93.66%	55.43%	91.18%
	DT	89.23%	89.23%	92.31%	57.15%	90.37%
	FDA	92.77%	92.77%	92.51%	47.09%	92.41%
	KNN	91.39%	91.39%	93.41%	53.47%	91.77%

Firstly, the accuracy rate of 91.12% for the NB classifier shows a marginal improvement over the baseline performance of the HMSA-IVECM System. This is a slight improvement in the accurate classification of cases. Furthermore, NB shows a higher precision of 93.49%, suggesting a better ability to correctly classify positive cases. However, the sensitivity/recall achieved 91.12% along with this improvement in accuracy. In addition, the specificity of the NB classifier shows a slight improvement to 53.19%, indicating a stronger capacity to accurately detect negative instances. With improvements in both precision and specificity. The SVM classifier achieves an accuracy of 90.44%. Additionally, the SVM classifier shows a higher precision of 93.66%, suggesting an improved capacity to correctly categorise positive cases. Furthermore, the specificity of the SVM classifier registers at 55.43%, indicating an improvement in its capacity to accurately detect TN cases. The SVM classifier performs well overall, showing improvements in specificity and precision but a little drop in sensitivity. The DT classifier, and at 89.23%, it performs close to baseline accuracy of the HMSA-IVECM System. In spite of this, the DT classifier shows a higher precision of 92.31%, suggesting a more advanced capacity to correctly categorise positive cases. Furthermore, the highest specificity is achieved by the DT classifier at 57.15%, suggesting that it can reliably detect TN cases. On the other hand, the accuracy of the FDA classifier has significantly improved to 92.77%, indicating a minor improvement in accurately identifying cases. Additionally, the specificity of the FDA classifier drops to 47.09%, indicating a decreased ability to accurately detect real negative cases. The FDA classifier yields mixed results overall, increasing accuracy but decreasing specificity. Finally, the KNN classifier has a higher precision of 93.41%, demonstrating an improved capacity to correctly categorise positive cases. Additionally, the specificity of the KNN classifier shows a minor improvement to 53.47%, indicating a stronger capacity to accurately classify abnormal cases. Overall, the KNN classifier performs well, with increased accuracy, precision, specificity and f1-sccore.

Table 10.4 presents the results of the evaluation of performance of the HMSA-IVECM System following the incorporation of the f-test feature selection model with various classifiers. The results show that the performance of each classifier varies somewhat across several performance indicators.

Table 10.4 Performance Evaluation of HMSA-IVECM System Employing the f-test Feature Selection Model with Various Classifiers

Different integrated contexts	Classifier	Performance Measure				
		Accuracy	Sensitivity/Recall	Precision	Specificity	F1-Score
SelfAttention_ModAlexNet training on SGDM		90.99%	90.99%	91.98%	51.99%	91.26%
Feature Selection (f-test)	**NB**	88.62%	88.62%	90.87%	50.19%	89.56%
	SVM	90.38%	90.38%	93.61%	54.40%	91.12%
	DT	88.39%	88.39%	92.87%	57.98%	90.02%
	FDA	92.54%	92.54%	92.66%	48.10%	92.34%
	KNN	93.55%	93.55%	93.87%	53.74%	93.31%

There is a noticeable drop in accuracy from the baseline, starting with the NB classifier, dropping from 90.99% to 88.62%. With 90.87% precision after feature selection, it is still comparatively high. Significantly, specificity has decreased from 51.99% to 50.19%, suggesting a possible loss in accurately categorising negative cases. The F1-Score holds at 89.56% despite these adjustments. The SVM classifier exhibits performance that is essentially in line with the baseline, with the exception of a minor drop in accuracy from 90.99% to 90.38%. Precision improves significantly to 93.61%, indicating enhanced discriminatory power. Furthermore, the specificity has improved marginally from the baseline to 54.40%. As a result, at 91.12%, the F1-Score is comparatively unchanged. With regard to the DT classifier, accuracy is comparable to that of NB, indicating a minor decline from the baseline to 88.39%. Similar to NB, sensitivity is still at 88.39%. Nevertheless, DT attains 92.87% precision after feature selection, which is greater than the baseline model. Most remarkably, the specificity of the DT classifier, which rises to 57.98%, improves significantly when compared to the baseline and other classifiers. With an accuracy of 92.54%, FDA outperforms the baseline model. With a precision of 92.66% and sensitivity of 92.54%, the performance is balanced. Specificity, which dropped to 48.10%, is noticeably lower than that of the baseline and other classifiers. Even yet, the F1-Score, at 92.34%, indicating the increase of normal cases classification over the classification of abnormal classes. Finally, with a significant improvement to 93.55%, the KNN classifier attains the greatest accuracy of all the classifiers. Furthermore, KNN achieves a high precision of 93.87% in post-feature selection. Additionally, specificity increased somewhat from the baseline to 53.74%.

As a result, the outstanding overall performance of the KNN classifier is reflected in the F1-Score, which is 93.31%. In summary, FDA and KNN demonstrate high accuracy and precision, despite the fact that the performance of each classifier varies across different criteria. Significant improvements in specificity show that DT is useful for accurately classifying negative cases.

The ensemble model is deployed intentionally following the proposed feature selection model, which demonstrates a logical strategy to improving prediction precision. As explained in the IVECM, its structure consists of two key phases that are highly linked: class weight assignment and classifier weight assignment. This phase is crucial because it establishes the relative weights assigned to various classes and classifiers, which in turn affects the decision-making process that follows. After these steps are carefully carried out, the ensemble model produces two important performance metrics, which are calculated meticulously for each classifier: the maximal voting F1-score measure and the maximal voting specificity measure. To ensure an accurate evaluation of the discriminative abilities of the model, the maximal voting specificity metric provides information regarding the ability of the model to correctly detect real negatives. Conversely, the maximal voting F1-score measure provides a thorough assessment of the predictive effectiveness of the model by encapsulating both accuracy and recall performance.

By combining the performance of each classifier with an intelligent combination to minimise the limitations of each one, these methods conclude in the development of a maximal voting stacking ensemble model. The objective of the ensemble model is to outperform any single classifier in terms of prediction by utilising the combined knowledge of several classifiers. Notably, the effectiveness of the ensemble model is emphasised by the results presented in Table 10.5, which provide conclusive proof of its predictive analytics capabilities.

A thorough examination of the performance of the ensemble model in multiple integrated contexts, as shown in Table 5, provides significant insights into how the performance of different feature selection techniques affect predictive power. All configurations are carefully analysed, considering important performance indicators like F1-score, accuracy, precision, sensitivity/recall, and highlight comparative analyses.

Starting with the ensemble configurations that are integrated with the mRMR feature selection approach, accuracy increases slightly to 91.76%, while precision, specificity, and F1-score all increase as well. Significantly, the specificity rises to 53.97%, indicating higher TN detection than the baseline configuration. Comparable performances are also shown by ensemble models that use the f-test and Chi-Square feature selection techniques, with accuracies of 91.69% and 91.12%, respectively. These configurations highlight the effectiveness of these feature selection strategies in improving the discriminative abilities of the model by demonstrating increases in accuracy, specificity, and F1-score. Slight differences can be seen when specificity is compared in detail. Chi-
Square achieves higher specificity, at 53.67%, closely followed by the f-test, at 53.52%.

Table 10.5 Performance Evaluation of the ensemble model

Different integrated contexts		Performance Measure				
		Accuracy	**Sensitivity/Recall**	**Precision**	**Specificity**	**F1-Score**
SelfAttention_ModAlexNet training on SGDM		90.99%	90.99%	91.98%	51.99%	91.26%
Ensemble for MRMR	**F1**	91.76%	91.76%	93.31%	53.42%	92.02%
	Specificity	91.30%	91.30%	93.15%	53.97%	91.72%
Ensemble for Chi-Square	**F1**	91.12%	91.12%	93.35%	53.10%	91.58%
	Specificity	91.07%	91.07%	93.39%	53.67%	91.57%
Ensemble for f-test	**F1**	91.69%	91.69%	93.37%	53.42%	91.99%
	Specificity	91.23%	91.23%	93.39%	53.52%	91.69%
Final Ensemble		90.13%	90.13%	92.77%	62.20%	91.03%

The combined strength provided by ensemble modelling is further illuminated through a thorough comparative study between the 'Final Ensemble' configuration and the performance of each classifier using each feature selection method. Even though individual configurations might perform better in some metrics—like accuracy or precision—the 'Final Ensemble' demonstrates a balanced performance across a number of parameters, highlighting its strength and adaptability. For example, the 'Final Ensemble' achieves an accuracy of 90.13%, whereas the MRMR ensemble model achieves a slightly higher accuracy of 91.76%. The 'Final Ensemble' performs

significantly better than the MRMR model in terms of specificity, though, with a specificity of 62.20% as opposed to 53.97% for the MRMR model. Comparable to the "Final Ensemble" in terms of accuracy, the specificity values of the Chi-Square and f-test ensemble models are noticeably lower. To be more precise, the f-test ensemble model obtains an accuracy of 91.69% and a specificity of 53.52%, whereas the Chi-Square ensemble model achieves an accuracy of 91.12% and 73.67%. The final Ensemble scores 91.03% and 92.77% in precision and recall, respectively, indicating a well-balanced performance. This comprehensive comparison highlights the thorough methodology of the 'Final Ensemble', which incorporates learnings from the diverse performance of various classifiers using multiple feature selection strategies. As a result, an extensive predictive model with exceptional overall predictive performance is generated.

The final ensemble model is most effective at prediction, especially when it comes to how well it performs in terms of specificity compared to the other models at the same accuracy level. This differentiation is evidence of the meticulous blending of various feature selection methods and classifier performances within the ensemble framework. The final ensemble model exhibits a surprising capacity to identify real negatives, as indicated by its better specificity measure, even if it achieves similar accuracy levels to other models. This resulted in a comprehensive predictive model that could provide reliable and precise predictions in a variety of scenarios, highlighting the skill of the ensemble model in reducing the shortcomings of individual classifiers and feature selection techniques. An achievement like this confirms the effectiveness of ensemble modelling and shows how it improves prediction precision in scenarios where specificity is crucial.

Table 10.6 Comparative Analysis of the Performance of the Five Systems

Systems	Performance Measure				
	Accuracy	Sensitivity/Recall	Precision	Specificity	F1-Score
DE System	89.60%	89.60%	87.17%	21.43%	88.34%
MA System	91.61%	91.61%	88.16%	23.91%	89.57%
FFS-EC	94.91%	94.91%	94.90%	43.07%	93.79%
MHMA System	90.99%	90.99%	91.98%	51.99%	91.26%
HMSA-IVECM System	90.13%	90.13%	92.77%	62.20%	91.03%

The comparative study in Table 10.6 of the performance of the five systems shows distinctive patterns regarding a number of performance indicators. It attains an accuracy of 89.60%, indicating a balanced sensitivity/recall of 89.60%, starting with DE System. Still, at 87.17% and 21.43%, respectively, precision and specificity are relatively low. By comparison, MA System has a somewhat greater precision but a higher accuracy of 91.61% with comparable sensitivity/recall. Nevertheless, DE System and MA System both display noticeably low specificity values, suggesting possible limits in accurately detecting real negatives. The accuracy of FFS-EC has significantly improved to 94.91%, and its precision, specificity, and sensitivity/recall measures are all balanced. This implies that, in comparison to DE and MA Systems, FFS-EC System performs more comprehensively. However, the specificity, at 43.07%, is still somewhat low despite the excellent accuracy and precision. MHMA System has a greater precision of 91.98% but an accuracy of 90.99%, close to the accuracy of the MA System. Even still, at 51.99%, its specificity is still very low. Finally, HMSA-IVECM System achieves the second highest precision of 92.77% and an accuracy of 90.13%. Furthermore, it outperforms all other systems in terms of specificity, achieving 62.20%. Even though the accuracy of the HMSA-IVECM System is marginally lower than that of MA System and MHMA System, it is far better at identifying real negatives.

Table 10.7 *Assessment of Performance Metrics and Statistical Significance of Enhancement via Cohen's d and T-test: HMSA-IVECM System Compared to DE, MA, FFS-EC, and MHMA Systems*

System	Cohen's d:		T-test (p)	
	Measure	Significance	Measure	Significance
DE System	0.4328	Medium	0.2651	No Significance
MA System	0.3615	Medium	0.3364	No Significance
FFS-EC System	0.050064	Small	0.84803	No Significance
MHMA System	0.11726	Small	0.44165	No Significance

A thorough analysis of performance measures and statistical significance related to the improvement that HMSA-IVECM System obtained over DE, MA, FFS-EC, and MHMA Systems is provided in Table 10.7. The evaluation uses the T-test

significance testing and Cohen's d significance measure to measure and validate the observed performance variations across different systems.

In comparison to DE System, HMSA-IVECM System has a notable Cohen's d value of 0.4328, indicating a medium impact size. This indicates a significant performance difference, but the corresponding T-test produces a p-value of 0.2651, indicating statistical insignificance. However, it is important to note that the impact size still indicates a significant improvement in HMSA-IVECM System over DE System. In comparison to HMSA-IVECM System, MA System shows an acceptable Cohen's d value of 0.3615, indicating a medium impact size. Although the impact size is significant, the T-test yields a non-significant p-value of 0.3364. Once more, the impact size indicates a significant improvement of HMSA-IVECM System over MA System.

In comparison with DE and MA Systems, HMSA-IVECM System exhibits a much lower Cohen's d value of 0.050064, suggesting that its improvement impact is smaller compared to FFS-EC. As a result, the corresponding T-test yields a high p-value of 0.84803. This comparison confirms that HMSA-IVECM System is superior to FFS-EC even with the small significance effect size. Compared to MHMA System, HMSA-IVECM System exhibits a Cohen's d value of 0.11726, indicating a smaller impact size. This improvement, nevertheless, the T-test results in a non-significant p-value of 0.44165. The effect size indicates that HMSA-IVECM System is significantly better than MHMA System even though there is no statistical significance.

In summary, the significant effect sizes consistently show that HMSA-IVECM System is superior to DE, MA, FFS-EC, and MHMA Systems even though statistical significance may not be reached consistently across comparisons. In conclusion, even if the T-test findings for DE, MA, FFS-EC, and MHMA Systems might not all reach conventional significance thresholds, each test offers important proof of the importance of the improvement of the HMSA-IVECM System over the corresponding comparative systems. The reason for superiority of the HMSA-IVECM System in this situation is supported by the constant trends of the comparisons towards significance.

To compare the system with other studies using the same dataset, the system was retrained for binary classification, implementing two scenarios: one that classifies images as either normal or abnormal, with the abnormal class including both benign and malignant cases, and another that classifies images as cancerous or non-cancerous, with non-cancerous encompassing both benign and normal cases. These scenarios were chosen to ensure a fair comparison with other studies, as some classify as either normal and abnormal, while others classify them as cancerous and non-cancerous.

Table 10.8 Performance Evaluation of Binary Classification Scenarios: Differentiating Normal vs. Abnormal (Scenario 1) and Cancerous vs. Non-Cancerous(Scenario 2)

Scenario	Accuracy	Sensitivity/Recall	Precision	Specificity	F1-Score
Scenario 1	94.53%	95.83%	98.20%	79.03%	97.00%
Scenario 2	93.81%	94.01%	99.60%	87.20%	96.72%

The results of the two binary classification scenarios is shown in Table 10.8. In scenario 1, images were classified as normal or abnormal, with both benign and malignant cases included in the abnormal class, and the system achieved a remarkable accuracy of 94.53%. This signifies the ability of the system in the classification of images across a range of abnormalities. On the other hand, Scenario 2 achieved a lower accuracy of 93.81%, where images were classified as cancerous or non-cancerous, with non-cancerous including both benign and normal classes. Moreover, Scenario 2 exhibits an impressive precision of 99.60%, while Scenario 1 achieves a slightly lower precision of 98.20%. Notably, Scenario 2 also achieves a higher specificity of 87.20% compared to Scenario 1.

The superior performance of Scenario 2 can be attributed to its clear distinction between normal/benign and malignant cases. By separating these classes, the classification system focuses more precisely on the identification of features specific to malignancy, leading to better classification performance. On the other hand, the combination of benign and malignant cases into a single class in Scenario 1 may have introduced ambiguity and make it more challenging for the classification model.

Numerous studies have investigated the classification of DBT scans using the BCS-DBT dataset. However, a common trend among these studies is the integration of

selected subsets from the BCS-DBT dataset with private DBT datasets, potentially introducing variability in the results. Furthermore, classification strategies varied among studies, with some classifying images as benign versus malignant, while others distinguished between normal versus abnormal cases, where "abnormal" included both benign and malignant classifications. Additional approaches included differentiating among normal, benign, and malignant categories, or classifying images as cancerous versus non-cancerous, with the non-cancerous category encompassing both benign and normal cases.

To address this challenge, multiple iterations of the system were conducted to ensure a fair comparison. When comparing the results with studies that integrated private datasets, the same data subset was used, and the classification approach was modified accordingly. The system was executed in three distinct configurations: first, for multi-class classification into normal, benign, and malignant categories; second, for a binary classification of normal versus abnormal, with "abnormal" encompassing both benign and malignant cases; and third, for the classification of cancerous versus non-cancerous instances, where "non-cancerous" included both benign and normal cases. Our system was tested in three previously mentioned configurations, and the results are presented and compared with previous work in Tables 10.9, 10.10, and 10.11.

Table 10.9 Results of Multi-Class Classification for Normal, Benign, and Malignant Classes

Author/Year	Accuracy	Sensitivity/Recall	Precision	Specificity	F1-Score
Our System (HMSA-IVECM)	90.13%	90.13%	92.77%	62.20%	91.03%
Nogay, Akinci, and Yilmaz, (2021)	75.00%				

Table 10.9 shows that our system outperformed the model by Nogay, Akinci, and Yilmaz (2021) in the multi-class classification, achieving 90.13% accuracy compared to their 75.00%. This improvement is due to the use of a more advanced architecture, specifically the multi-head self-attention mechanism and ensemble classification model, which enhanced feature extraction and combined the strengths of multiple classifiers. In contrast, Nogay et al. (2021) relied on pre-trained DCNNs

with transfer learning, which may have been less suited to capturing the complexities of DBT images. As a result, our system demonstrated superior performance in multi-class classification tasks.

Table 10.10 Results of Classification for Cancerous versus Non-Cancerous Classes

Author/Year	Accuracy	Sensitivity/Recall	Precision	Specificity	F1-Score	AUC
Our System (HMSA-IVECM)	93.81%	94.01%	99.60%	87.20%	96.72%	0.91
Tardy and Mateus (2021)						0.73
Nogay, Akinci, and Yilmaz, (2021)	86.00%					
Bai et al. (2022)	84.00%	84.00%	86.00%		83.00%	
Adhikesaven et al. (2022)	97.25%					
Bai et al. (2022)	92.00%	93.00%	91.00%	91.00%	92.00%	

In comparing the results from Table 10.10 for the classification of cancerous versus non-cancerous cases, the performance metrics of various systems reflect different strengths. Our system achieved an accuracy of 93.81%, a sensitivity of 94.01%, a high precision of 99.60%, and a specificity of 87.20%, with an F1-score of 96.72% and an AUC of 0.91. These metrics indicate that our multi-head self-attention and ensemble classification model is highly effective, particularly in precision and F1-score, demonstrating its strength in accurately identifying cancerous cases while maintaining a balanced ability to distinguish non-cancerous instances.

In contrast, the study by Tardy and Mateus (2021) reported an AUC of 0.73. Their approach, while novel with a deep multiple-instance-based learning model and trainable summarization for DBT, had lower overall performance, likely due to the complexity of processing DBT images and their reliance on a private multi-vendor dataset, which may have introduced variability in results.

Nogay, Akinci, and Yilmaz (2021) reported an accuracy of 86.00%. Their use of pre-trained DCNNs with transfer learning for binary classification proved effective but did not reach the precision or specificity levels achieved by our system. Their methodology, which leveraged classical DCNN architectures like AlexNet and ResNet, lacked the advanced attention mechanisms and feature extraction techniques incorporated in our model.

Bai et al. (2022) demonstrated accuracies of 84.00% and 92.00%, depending on the experiment, with sensitivity reaching 93.00%. Their approach used graph convolutional networks, which proved useful but did not match the precision or specificity of our system, likely due to differences in how features were integrated and selected for final classification.

In the study by Bai et al. (2022), the feature fusion Siamese network achieved strong performance metrics, including an accuracy of 92.00%, sensitivity of 93.00%, precision of 91.00%, specificity of 91.00%, and an F1-score of 92.00%. These results can be attributed to their innovative approach of comparing both current and prior mammogram images, mimicking the workflow of radiologists in detecting subtle changes over time. The Siamese network model, combined with their distance learning technique, allowed for better feature extraction and comparison between images, improving classification.

Lastly, Adhikesaven et al. (2022) achieved the highest accuracy of 97.25%, using a CNN specifically designed for early breast cancer detection. Despite this impressive accuracy, the lack of reported precision and specificity makes it difficult to directly compare with our model, which balances high precision and specificity while maintaining strong overall performance.

Table 10.11 Results of Classification for Normal versus Abnormal Classes

Author/Year	Accuracy	Sensitivity/Recall	Precision	Specificity	F1-Score	AUC
Our System (HMSA-IVECM)	94.53%	95.83%	98.20%	79.03%	97.00%	0.87
Du et al. (2024)		84.62%		84.15%		0.92
Fogleman, Otsap, and Cho (2021)	94.90%					

In comparing the results from Table 10.11 for the classification of normal versus abnormal cases, our system achieved a high accuracy of 94.53%, with a sensitivity of 95.83%, precision of 98.20%, specificity of 79.03%, an F1-score of 97.00%, and an AUC of 0.87. In contrast, Du et al. (2024) reported an accuracy of 84.62% and an AUC of 0.92. While Du et al. employed a novel self-supervised initialization and fine-tuning strategy (SIFT-DBT) for imbalanced data, their lower accuracy compared to

our system may be attributed to the challenges of addressing data imbalance using their patch-level multi-instance learning approach.

Similarly, Fogleman, Otsap, and Cho (2021) achieved an accuracy of 94.90%, which is slightly higher than ours. Their system utilized transfer learning with partial Inception v3 architecture, which helped achieve strong results. However, the other performance metrics of their model, such as specificity and precision, were not reported, limiting the depth of comparison. Overall, the balance between precision, specificity, and F1-score in our system highlights its robustness in classifying normal versus abnormal cases, surpassing the approach of Du et al. while performing comparably in terms of accuracy to Fogleman et al.

Only a limited number of researchers have developed systems utilizing solely the BCS-DBT subset without incorporating private datasets. In comparisons with studies that exclusively used the BCS-DBT dataset, the same data utilized by those researchers was adopted, and the system was adjusted to align with their classification models, which primarily focused on distinguishing between benign and malignant cases. Several scenarios were considered to ensure that the comparisons and validations were conducted fairly. The results are presented in Tables 10.12, 10.3, and 10.4.

Table 10.12 Comparison of Classification Results for Benign vs Malignant Cases Using the BCS-DBT Dataset only (Scenario 1)

Author/Year	Accuracy	Sensitivity/Recall	Precision	Specificity	F1-Score
Our System (HMSA-IVECM)	91.09%	92.74%	88.49%	89.73%	90.45%
Hassan et al. (2022)	80.43%				
Hassan et al. (2024)	85.00%	90.00%	84.10%		86.90%

In comparing the results for Table 10.12, our system achieved an accuracy of 91.09%, with a sensitivity of 92.74%, precision of 88.49%, specificity of 89.73%, and an F1-score of 90.45%. These performance metrics indicate that our system effectively classified benign and malignant cases, showing a strong balance between sensitivity and specificity. In contrast, Hassan et al. (2022) reported a lower accuracy of 80.43%. Their approach focused on deep learning-based radiomics with SVM

classification, which, while effective, was likely limited by the ability of their model to extract comprehensive features from the DBT dataset. The reliance on a smaller training dataset and the application of traditional machine learning techniques like SVM may have restricted the performance of the model.

In their subsequent work, Hassan et al. (2024) achieved improved accuracy at 85.00%, with a sensitivity of 90.00%, precision of 84.10%, and an F1-score of 86.90%. The improvement can be attributed to the introduction of image quality-aware features and tumour texture descriptors, which enhanced feature extraction and classification. However, the system still underperformed compared to ours, likely due to the lack of an ensemble model and advanced attention mechanisms that are central to the HMSA-IVECM system. Overall, our system outperformed both versions of the models by Hassan.

Table 10.13 Comparison of Classification Results for Benign vs Malignant Cases Using the BCS-DBT Dataset only (Scenario 2)

Author/Year	Accuracy	Sensitivity/Recall	Precision	Specificity	F1-Score
Our System (HMSA-IVECM)	91.24%	91.35%	89.69%	91.14%	90.40%
Farangis Sajadi Moghadam and Rashidi (2023)	88.67%	77.12%		75.11%	
Farangis Sajadi Moghadam and Rashidi (2024)	78.51%	82.78%		75.19%	

In comparing the results for Table 10.13, our system achieved an accuracy of 91.24%, with a sensitivity of 91.35%, precision of 89.69%, specificity of 91.14%, and an F1-score of 90.40%. These strong metrics reflect the robust performance of our system, which combines advanced feature extraction and classification techniques to effectively distinguish between benign and malignant cases. In contrast, Farangis Sajadi Moghadam and Rashidi (2023) achieved a lower accuracy of 88.67%, with a sensitivity of 77.12% and specificity of 75.11%. Their method, which utilized radiomic-based feature extraction and Quadratic Discriminant Analysis (QDA), was effective but fell short in sensitivity and precision. This could be due to the limitations in feature extraction and model choice, which may have struggled to capture the full complexity of the DBT images.

In a later study, Farangis Sajadi Moghadam and Rashidi (2024) reported an even lower accuracy of 78.51%, with a sensitivity of 82.78% and specificity of 75.19%.

Despite introducing a novel feature extraction method based on DCT-DOST features, the performance of their model remained limited. The relatively lower precision and specificity suggest that their system may have overfitted the training data, particularly with the use of smaller sample sizes, which could have led to poor generalization on the test data. Overall, our system outperformed both studies by Farangis Sajadi Moghadam and Rashidi, particularly in terms of accuracy, sensitivity, and specificity.

Table 10.14 Comparison of Classification Results for Benign vs Malignant Cases Using the BCS-DBT Dataset only (Scenario 3)

Author/Year	Accuracy	Sensitivity/Recall	Precision	Specificity	F1-Score
Our System (HMSA-IVECM)	90.34%	88.24%	90.12%	92.02%	89.09%
J. Mendes et al. (2023)	93.20%	92.00%	94.00%	94.00%	94.00%

In comparing the results from Table 10.14, our system achieved an accuracy of 90.34%, with a sensitivity of 88.24%, precision of 90.12%, specificity of 92.02%, and an F1-score of 89.09%. In contrast, the study by J. Mendes et al. (2023) reported higher performance metrics with an accuracy of 93.20%, sensitivity of 92.00%, precision of 94.00%, specificity of 94.00%, and an F1-score of 94.00%. The higher performance of the model by Mendes et al. can be attributed to their unique approach of utilizing single-slice DBT classification. This approach effectively addresses the tissue overlap problem found in 2D mammography, allowing their model to perform at a higher level. By using nine slices from each DBT volume and applying extensive data augmentation techniques, Mendes et al. increased the variability and quantity of training data, which likely contributed to the enhanced performance of their CNN model. Additionally, the careful extraction of regions of interest (ROIs) and preprocessing further boosted the accuracy of the model in differentiating between benign and malignant cases. In comparison, while our model still performed well, the results suggest that further optimization, perhaps through similar ROI selection techniques or slice-based analysis, could narrow the gap between the two systems.

10.4 Summary

This chapter was devoted to improving the classification of DBT data using MHMA System, which combined Mod_AlexNet with a Multi-Head Attention model. The DBT classification issues were successfully handled by this method, which also showed enhanced data extraction and performance. HMSA-IVECM System was developed in response to observations from FFS-EC that highlighted the value of feature fusion and ensemble approaches. Combining improvements from FFS-EC and MHMA Systems, HMSA-IVECM System introduces an ensemble model to address challenges with different breast densities, sizes and abnormal cases classification. Particularly, HMSA-IVECM System outperforms traditional pre-trained DCNN models and offers a significant improvement in multi-class classification. Image augmentation, feature extraction using Mod_AlexNet and attention models, and feature fusion using HOG descriptors are all part of its architecture. HMSA-IVECM System incorporates strong feature selection techniques, as well as ensemble models and high-performing classifiers, to improve classification. Class and classifier weights are introduced in HMSA-IVECM System, which offers improvements over conventional models in terms of decision-making and prediction precision, especially in a stacking ensemble model. Overall, the development of HMSA-IVECM System represents an important advancement in DBT classification techniques, offering enhanced efficacy.

Chapter 11 Conclusion and Further work

This chapter provides concluding observations and discusses potential directions for future work. The first section summarises the results and key conclusions derived from this thesis, while the second section outlines possible directions for future research and development.

11.1 Conclusion

Breast cancer remains the most prevalent cancer among women worldwide, with the highest mortality rate. Early detection significantly improves patient outcomes by enabling timely treatment, which increases survival rates. This thesis aimed to accurately distinguish between normal, benign, and malignant tomosynthesis scans, regardless of variations in breast size, breast density, or tumour size. A key focus of the research was enhancing the ability to differentiate between benign and malignant tumours, a particularly challenging task due to the small differences in appearance within tomosynthesis scans. In this thesis, novel automated DBT multi-class classification systems were presented, which integrate both ML and DL techniques to classify DBT scans. The BCS-DBT dataset was utilized to evaluate the performance of the systems, and several key performance metrics were calculated. Five systems were developed and evaluated in this research, as summarised below. In Chapter 6, the DE System evaluated the effectiveness of six state-of-the-art deep learning models for feature extraction, coupled to a classification stage which used an SVM classifier. Among these, AlexNet achieved the highest performance compared to the other deep learning models; however, further enhancements were required, particularly in specificity, which was limited, with AlexNet achieving the highest specificity at only 21.43%. Building on this, the Mod_AlexNet system, representing first knowledge contribution of the thesis, was developed as the second system in this thesis and was presented in Chapter 7. Mod_AlexNet was designed and developed, in Chapter 7, to address the challenge of low specificity and to enhance multiclass classification accuracy by extracting features that better distinguish between classes, improving class separability regardless of variations in breast density, breast size, and tumour size. Mod_AlexNet is a modified version of the original AlexNet architecture, with the addition of max-pooling and batch normalization layers, which resulted in enhanced performance across all

performance measures. When compared to the DE System, presented in Chapter 6, Mod_AlexNet achieved a 2.24% improvement in accuracy and 11.59% improvement in specificity.

Despite the improvements, the classification of benign and malignant remained challenging. In Chapter 8, building on these findings, feature fusion and selection models were explored and utilised to address the multi-class classification challenge. By concatenating features from deep learning models and HOG descriptors, feature fusion allows system to capture broader spectrum of image features, improving class separability. Feature selection further eliminates irrelevant features, which can reduce classification performance. The FFS-EC System, the second contribution, was developed. This system integrates feature fusion and selection with a majority voting ensemble classifier, utilising ResNet-50, SqueezeNet, and Mod_AlexNet for feature extraction. The FFS-EC System achieved a 94.91% accuracy, a 5.93% improvement rate over the baseline, and a 43.07% specificity, compared to just 21.43% in the baseline. However, the challenge of accurately classifying the abnormal class, as reflected in the specificity measure, persisted.

In Chapter 9, MHMA System is introduced to address the challenge of accurately classifying tumours as benign or malignant due to the small differences between them. MHMA System incorporates a novel Multi-Head Self-Attention model with Mod_AlexNet, representing the third contribution of this thesis. The attention mechanism enhances the ability of the model to focus on relevant image regions, improving class discrimination. The MHMA System demonstrated a specificity of 51.99% and an F1-score of 91.28%, significantly surpassing the 21.43% specificity and 88.34% F1-score of the baseline. These results underscore the importance of the attention model in enhancing classification performance, especially in benign and malignant classification.

Finally, in Chapter 10, after analysing the results from the FFS-EC and MHMA Systems, HMSA-FFS-IVECM System was developed, which integrates the strengths of these systems. This final system employs various feature selection techniques post MHMA feature extraction and inputs the selected features into the IVECM ensemble classifier, which represents the final contribution of this thesis. This ensemble classifier, which incorporates class and classifier weights, maximized the

classification performance. The system successfully addressed key objectives, such as handling the non-linearity of class boundaries, enhancing class separability, and reducing effect of class imbalance. The system demonstrated significant improvements across all metrics, achieving 90.13% accuracy, 92.77% precision, 62.20% specificity, and a 91.03% F1-score. These results represent a significant enhancement over the baseline, particularly in specificity, which achieved an improvement of 40.77% compared to the DE System.

When comparing to previous work using the same dataset, the HMSA-FFS-IVECM System not only outperformed existing multi-class classifiers (normal, benign, and malignant classes) but also excelled in binary classification systems (such as distinguishing between normal and abnormal, benign and malignant, or cancerous and non-cancerous). The system achieved outstanding results, surpassing the performance of other methodologies in the field, as demonstrated in Tables 10.9 to 10.14 in Section 10.3.

In conclusion, this dissertation made substantial contributions to the development of an automated DBT multi-class classification system, which can assist radiologists in the diagnosis process, leading to earlier detection and improved patient outcomes. As the field of DBT classification continues to evolve, the insights gained from this research will serve as a foundational element for future intelligent CAD Systems aimed at enhancing the accuracy and reliability of DBT scan classification.

11.2 Future Work

The work presented in this dissertation has made significant strides in the development of an automated DBT multi-class classification system, aimed at reducing errors made by radiologists while diagnosing scans. However, several aspects for future research remain open to enhance the classification performance. The following future directions are suggested:

- Extracting additional deep features from the DBT scans and integrate them into the current systems to capture more complex information, such as tissue density patterns and small texture variations, which can help in better discrimination between benign and malignant tumours, which enhances the specificity measure.

- Fusing multi-modality imaging data for enhancement of performance. The fusion of DBT with other imaging modalities, such as MRI or US. MRI offers more detailed view of soft tissue and ultrasound is able to differentiate between solid and fluid-filled lumps. This multi-modality fusion could provide a more detailed diagnostic perspective, thereby improving the discrimination between different classes.
- Investigating the use of Generative Adversarial Networks (GANs) to augment the training dataset, especially minority class to overcome class imbalance and improve overall classification performance.
- Building upon the Multi-Head Self-Attention model to enhance the specificity, particularly in challenging cases where distinguishing between classes is difficult.

References

Adhikesaven, S., Kapoor, A., Khowaja, A.S., Li, V., Sabhanayakam, K. and McMahan, L. (2022). Predicting the Instance of Breast Cancer within Patients using a Convolutional Neural Network. *Journal of Emerging Investigators*. [online] doi:https://doi.org/10.59720/22-061.

Ajay Kumar Visvkarma, Khushwant Sehra, Chanchal Saraswat, Radhapiyari Laishram, Malik, A., Sharma, S.K., Kumar, S., Rawal, D.S., Vinayak, S. and Saxena, M. (2022). Impact of Gamma Radiations on Static, Pulsed *I–V*, and RF Performance Parameters of AlGaN/GaN HEMT. *IEEE Transactions on Electron Devices*, 69(5), pp.2299–2306. doi:https://doi.org/10.1109/ted.2022.3161402.

Ali, E.A. and Adel, L. (2019). Study of the role of digital breast tomosynthesis over digital mammography in the assessment of BIRADS 3 breast lesions. Egyptian Journal of Radiology and Nuclear Medicine, 50(1).

Alom, M.Z., Taha, T.M., Yakopcic, C., Westberg, S., Hasan, M., Esesn, B.C., Awwal, A.A., Asari, V.K. (2018). The history began from Alexnet: A comprehensive survey on deep learning approaches. arXiv, arXiv:1803.01164.

Alshamrani, K., Alshamrani, H.A., Alqahtani, F.F. and Almutairi, B.S. (2022). Enhancement of Mammographic Images Using Histogram-Based Techniques for Their Classification Using CNN. *Sensors*, 23(1), p.235. doi:https://doi.org/10.3390/s23010235.

Alsharman, N. and Jawarneh, I. (2020). GoogleNet CNN Neural Network towards Chest CT-Coronavirus Medical Image Classification. *Journal of Computer Science*, 16(5), pp.620–625. doi:10.3844/jcssp.2020.620.625.

Antonios Tragoudaras, Pavlos Stoikos, Konstantinos Fanaras, Athanasios Tziouvaras, Floros, G., Dimitriou, G., Kostas Kolomvatsos and Stamoulis, G. (2022). Design Space Exploration of a Sparse MobileNetV2 Using High-Level Synthesis and Sparse Matrix Techniques on FPGAs. *Sensors*, 22(12), pp.4318–4318. doi:https://doi.org/10.3390/s22124318.

Arian, A., Dinas, K., Pratilas, G.C. and Alipour, S. (2022). The Breast Imaging-Reporting and Data System (BI-RADS) Made Easy. *Iranian Journal of Radiology*, 19(1). doi:https://doi.org/10.5812/iranjradiol-121155.

Arnold, M., Morgan, E., Rumgay, H., Mafra, A., Singh, D., Laversanne, M., Vignat, J., Gralow, J.R., Cardoso, F., Siesling, S. and Soerjomataram, I. (2022). Current and future burden of breast cancer: Global statistics for 2020 and 2040. *The Breast*, 66(66). doi:https://doi.org/10.1016/j.breast.2022.08.010.

Attallah, O. (2021). MB-AI-His: Histopathological Diagnosis of Pediatric Medulloblastoma and its Subtypes via AI. *Diagnostics*, 11(2), p.359. doi:https://doi.org/10.3390/diagnostics11020359.

Awais, M., Iqbal, Md.T.B. and Bae, S.-H. (2020). Revisiting Internal Covariate Shift for Batch Normalization. *IEEE Transactions on Neural Networks and Learning Systems*, pp.1–11. doi:https://doi.org/10.1109/tnnls.2020.3026784.

Bai, J., Jin, A., Jin, A., Wang, T., Yang, C. and Nabavi, S. (2022). Applying graph convolution neural network in digital breast tomosynthesis for cancer classification. *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. doi:10.1145/3535508.3545549.

Bai, J., Jin, A., Wang, T., Yang, C. and Nabavi, S. (2022). Feature fusion Siamese network for breast cancer detection comparing current and prior mammograms. *Medical Physics*, 49(6), pp.3654–3669. doi:https://doi.org/10.1002/mp.15598.

Bandyopadhyay, S.K. (2010). Survey on Segmentation Methods for Locating Masses in a Mammogram Image. International Journal of Computer Applications, 9(11), pp.25–28.

Bevilacqua, V., Brunetti, A., Guerriero, A., Trotta, G.F., Telegrafo, M. and Moschetta, M. (2019). A performance comparison between shallow and deeper neural networks supervised classification of tomosynthesis breast lesions images. *Cognitive Systems Research*, 53, pp.3–19. doi:https://doi.org/10.1016/j.cogsys.2018.04.011.

Bistoni, G. and Farhadi, J. (2015). Anatomy and physiology of the breast. *Plastic and reconstructive surgery*, pp.477–485. doi:https://doi.org/10.1002/9781118655412.ch37.

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R.L., Soerjomataram, I. and Jemal, A. (2024). Global Cancer Statistics 2022: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A cancer journal for clinicians*, 74(3), pp.229–263. doi:https://doi.org/10.3322/caac.21834.

Buda, M., Saha, A., Walsh, R., Ghate, S., Li, N., Swiecicki, A., Lo, J. Y., Yang, J., & Mazurowski, M. (2020). Breast Cancer Screening – Digital Breast Tomosynthesis (Breast-Cancer-Screening-DBT) [Data set]. The Cancer Imaging Archive. https://doi.org/10.7937/E4WT-CD02

Buda, M., Saha, A., Walsh, R., Ghate, S., Li, N., Swiecicki, A., Lo, J.Y. and Mazurowski, M.A. (2021). A Data Set and Deep Learning Algorithm for the Detection of Masses and Architectural Distortions in Digital Breast Tomosynthesis Images. *JAMA Network Open*, 4(8), p.e2119100. doi:https://doi.org/10.1001/jamanetworkopen.2021.19100.

Bui Hai Phong, Thang Manh Hoang and Le, T.-L. (2020). A Hybrid Method for Mathematical Expression Detection in Scientific Document Images. *IEEE Access*, 8, pp.83663–83684. doi:https://doi.org/10.1109/access.2020.2992067.

Chan, H.-P., Samala, R.K. and Hadjiiski, L.M. (2020). CAD and AI for breast cancer—recent development and challenges. *The British Journal of Radiology*, 93(1108), p.20190580. doi:https://doi.org/10.1259/bjr.20190580.

Chen, Y., Jiang, H., Li, C., Jia, X. and Ghamisi, P. (2016). Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks. *IEEE Transactions on Geoscience and Remote Sensing*, [online] 54(10), pp.6232–6251. doi:https://doi.org/10.1109/tgrs.2016.2584107.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L. and Prior, F. (2013). The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information

Repository. *Journal of Digital Imaging*, 26(6), pp.1045–1057. doi:https://doi.org/10.1007/s10278-013-9622-7.

Conant, E.F. (2014). Clinical Implementation of Digital Breast Tomosynthesis. Radiologic Clinics of North America,52(3), pp.499–518. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4128330/

Cui, C., Li, L., Cai, H., et al. (2021) The Chinese Mammography Database (CMMD): an online mammography database with biopsy confirmed types for machine diagnosis of breast. 2021.

Dalal, N. and Triggs, B., 2005, June. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)* (Vol. 1, pp. 886-893). Ieee.

Daly C, Puckett Y. New Breast Mass. In: StatPearls. StatPearls Publishing, Treasure Island (FL); 2023. PMID: 32809592.

DBTex Challenge, SPIE-AAPM-NCI DAIR Digital Breast Tomosynthesis Lesion Detection Challenge. Available online: https://www.aapm.org/ GrandChallenge/DBTex2/ (accessed on 24 January 2023).

Demetri-Lewis, A., Slanetz, P.J. and Eisenberg, R.L. (2012). Breast Calcifications: The Focal Group. *American Journal of Roentgenology*, 198(4), pp.W325–W343. doi:https://doi.org/10.2214/ajr.10.5732.

Du, Y., Hooley, R.J., Lewin, J. and Dvornek, Nicha C (2024). SIFT-DBT: Self-supervised Initialization and Fine-Tuning for Imbalanced Digital Breast Tomosynthesis Image Classification. *arXiv (Cornell University)*. doi:https://doi.org/10.48550/arxiv.2403.13148.

Erdogan, K. and Yilmaz, N. (2014). Shifting Colors to Overcome not Realizing Objects Problem due to Color Vision Deficiency. pp.11–14. doi:https://doi.org/10.15224/978-1-63248-034-7-27.

Fan, M., Zheng, H., Zheng, S., You, C., Gu, Y., Gao, X., Peng, W. and Li, L. (2020). Mass Detection and Segmentation in Digital Breast Tomosynthesis Using 3D-Mask

Region-Based Convolutional Neural Network: A Comparative Analysis. *Frontiers in Molecular Biosciences*, 7. doi:10.3389/fmolb.2020.599333.

Farangis Sajadi Moghadam and Rashidi, S. (2023). Classification of benign and malignant tumors in Digital Breast Tomosynthesis images using Radiomic-based methods. doi:https://doi.org/10.1109/iccke60553.2023.10326283.

Farangis Sajadi Moghadam and Rashidi, S. (2024). Novel feature extraction based on DCT-DOST features for classification of Digital Breast Tomosynthesis images into benign and malignant tumors. *Research Square (Research Square)*. doi:https://doi.org/10.21203/rs.3.rs-3931625/v1.

Faruk, T., Islam, M.K., Arefin, S. and Haq, M.Z. (2015). The Journey of Elastography: Background, Current Status, and Future Possibilities in Breast Cancer Diagnosis. Clinical Breast Cancer, 15(5), pp.313–324

Fda.gov. (2018). DBT Accreditation: It's Here. FDA. [online] Available at: https://www.fda.gov/radiation-emitting-products/mqsa-insights/dbt-accreditation-its-here.

Ferguson, M., Ak, R., Lee, Y.T.T. and Law, K.H., 2017, December. Automatic localization of casting defects with convolutional neural networks. In *2017 IEEE international conference on big data (big data)* (pp. 1726-1735). IEEE.

Ferreira, A.J. and Figueiredo, M.A. (2012). Boosting algorithms: A review of methods, theory, and applications. *Ensemble machine learning: Methods and applications*, pp.35-85.

Fleyeh, H., 2008. *Traffic and road sign recognition* (Doctoral dissertation), Dalarna University, School of Technology and Business Studies, Computer Engineering.

Fleyeh, H. and Roch, J. (2013). Benchmark Evaluation of HOG Descriptors as Features for Classification of Traffic Signs. *International Journal for Traffic and Transport Engineering*, 3(4), pp.448–464. doi:https://doi.org/10.7708/ijtte.2013.3(4).08.

Fogleman, S., Otsap, J. and Cho, S. (2021). Clinical Diagnosis Support with Convolutional Neural Network by Transfer Learning. *SMU Data Science Review*, [online] 5(3). Available at: https://scholar.smu.edu/datasciencereview/vol5/iss3/2 [Accessed 29 Nov. 2023].

Foody, G.M. and Mathur, A. (2004). A relative evaluation of multiclass image classification by support vector machines. *IEEE Transactions on geoscience and remote sensing*, *42*(6), pp.1335-1343.

Fotin, S.V., Yin, Y., Haldankar, H., Hoffmeister, J.W. and Periaswamy, S. (2016). Detection of soft tissue densities from digital breast tomosynthesis: comparison of conventional and deep learning approaches. *Medical Imaging 2016: Computer-Aided Diagnosis*. doi:10.1117/12.2217045.

Ganaie, M.A., Hu, M., Malik, A.K., Tanveer, M. and Suganthan, P.N. (2022). Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, *115*, p.105151.

Gao, M., Fessler, J.A. and Chan, H.-P. (2021). Deep Convolutional Neural Network With Adversarial Training for Denoising Digital Breast Tomosynthesis Images. *IEEE Transactions on Medical Imaging*, 40(7), pp.1805–1816. doi:https://doi.org/10.1109/tmi.2021.3066896.

Gao, M., Samala, R.K., Fessler, J.A. and Heang Ping Chan (2020). Deep convolutional neural network denoising for digital breast tomosynthesis reconstruction. *Medical Imaging 2020: Physics of Medical Imaging*. doi:https://doi.org/10.1117/12.2549361.

Ghojogh, B., Samad, M., Mashhadi, S., Kapoor, T., Ali, W., Karray, F. and Crowley, M. (2019). *Feature Selection and Feature Extraction in Pattern Analysis: A Literature Review*. [online] Available at: https://arxiv.org/pdf/1905.02845.pdf [Accessed 5 Mar. 2020].

Giaquinto, A.N., Sung, H., Miller, K.D., Kramer, J.L., Newman, L.A., Minihan, A., Jemal, A. and Siegel, R.L. (2022). Breast Cancer Statistics, 2022. *CA: A Cancer Journal for Clinicians*, 72(6). doi:https://doi.org/10.3322/caac.21754.

Goddi, A., Bonardi, M. and Alessi, S. (2012). Breast elastography: A literature review. Journal of Ultrasound, 15(3), pp.192–198.

Grandini, M., Bagli, E. and Visani, G., 2020. Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*.

Guido, R., Ferrisi, S., Lofaro, D. and Conforti, D. (2024). An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review. *Information*, [online] 15(4), p.235. doi:https://doi.org/10.3390/info15040235.

Guo, Y.R., Bai, Y.Q., Li, C.N., Bai, L. and Shao, Y.H., 2022. Two-dimensional Bhattacharyya bound linear discriminant analysis with its applications. *Applied Intelligence*, *52*(8), pp.8793-8809.

Guo, Z., Xie, J., Wan, Y., Zhang, M., Qiao, L., Yu, J., Chen, S., Li, B. and Yao, Y. (2022). A review of the current state of the computer-aided diagnosis (CAD) systems for breast cancer diagnosis. *Open Life Sciences*, [online] 17(1), pp.1600–1611. doi:https://doi.org/10.1515/biol-2022-0517.

Hanchuan Peng, Fuhui Long and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8), pp.1226–1238. doi:https://doi.org/10.1109/tpami.2005.159.

Harbeck, N., Penault-Llorca, F., Cortes, J., Gnant, M., Houssami, N., Poortmans, P., Ruddy, K., Tsang, J. and Cardoso, F. (2019). Breast cancer. *Nature Reviews Disease Primers*, 5(1). doi:https://doi.org/10.1038/s41572-019-0111-2.

Hassan, L., Abdel-Nasser, M., Saleh, A. and Puig, D. (2022). Breast Tumor Classification in Digital Tomosynthesis Based on Deep Learning Radiomics. *Frontiers in Artificial Intelligence and Applications*. doi:10.3233/faia220348.

Hassan, L., Abdel-Nasser, M., Saleh, A. and Puig, D. (2024). Classifying Breast Tumors in Digital Tomosynthesis by Combining Image Quality-Aware Features and Tumor Texture Descriptors. *Machine Learning and Knowledge Extraction*, [online] 6(1), pp.619–641. doi:https://doi.org/10.3390/make6010029.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.770–778. doi:10.1109/cvpr.2016.90.

Hela, B., Hela, M., Kamel, H., Sana, B. and Najla, M. (2013). Breast cancer detection: A review on mammograms analysis techniques. 10th International Multi-Conferences on Systems, Signals & Devices 2013 (SSD13).

Helvie, M.A. (2010). Digital Mammography Imaging: Breast Tomosynthesis and Advanced Applications. Radiologic clinics of North America, [online] 48(5), pp.917–929. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3118307/

Hooley, R.J., Scoutt, L.M. and Philpotts, L.E. (2013). Breast Ultrasonography: State of the Art. *Radiology*, 268(3), pp.642–659. doi:https://doi.org/10.1148/radiol.13121606.

Hossain, M.B., Nishikawa, R.M. and Lee, J. (2022). Developing breast lesion detection algorithms for digital breast tomosynthesis: Leveraging false positive findings. *Medical Physics*. doi:10.1002/mp.15883.

Huang, G., Liu, Z. and Weinberger, Kilian Q (2016). *Densely Connected Convolutional Networks*. [online] arXiv.org. Available at: https://arxiv.org/abs/1608.06993.

Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J. and Keutzer, K. (2016). SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and . *arXiv:1602.07360 [cs]*. [online] Available at: https://arxiv.org/abs/1602.07360.

IARC Working Group on the Evaluation of Cancer-Preventive Interventions. Breast cancer screening. Lyon (FR): International Agency for Research on Cancer. (2016). Screening Techniques. Available at: https://www.ncbi.nlm.nih.gov/books/NBK546557/

Iranmakani, S., Mortezazadeh, T., Sajadian, F., Ghaziani, M.F., Ghafari, A., Khezerloo, D. and Musa, A.E. (2020). A review of various modalities in breast imaging: technical aspects and clinical outcomes. *Egyptian Journal of Radiology and Nuclear Medicine*, 51(1). doi:https://doi.org/10.1186/s43055-020-00175-5.

Khvostikov, A., Karim Aderghal, Benois-Pineau, J., Krylov, A.S. and Gwénaëlle Catheline (2018). 3D CNN-based classification using sMRI and MD-DTI images for Alzheimer disease studies. *arXiv (Cornell University)*. doi:https://doi.org/10.48550/arxiv.1801.05968.

Kikinis, R. and Knutsson, H. (2000). Adaptive Image Filtering. *Elsevier eBooks*, pp.19–31. doi:https://doi.org/10.1016/b978-012077790-7/50005-9.

Kim, D.H., Kim, S.T., and Ro, Y.M. (2016). Latent feature representation with 3-D multi-view deep convolutional neural network for bilateral analysis in digital breast tomosynthesis. *In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 927–931. doi:10.1109/ICASSP.2016.7471811.

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), pp.84–90. doi:10.1145/3065386.

Kumar, A. (2022). *Support Vector Machine (SVM) Python Example*. [online] Data Analytics. Available at: https://vitalflux.com/classification-model-svm-classifier-python-example/.

Leiva, R.G., Anta, A.F., Mancuso, V. and Casari, P. (2019). A novel hyperparameter-free approach to decision tree construction that avoids overfitting by design. *IEEE Access*, *7*, pp.99978-99987.

Leung, K.M., 2007. k-Nearest Neighbor algorithm for classification. *Polytechnic University Department of Computer Science/Finance and Risk Engineering*.

Li, X., Qin, G., He, Q., Sun, L., Zeng, H., He, Z., Chen, W., Zhen, X. and Zhou, L. (2020). Digital breast tomosynthesis versus digital mammography: integration of image modalities enhances deep learning-based breast mass classification. *European Radiology*, 30(2), pp.778–788. doi:10.1007/s00330-019-06457-5.

Liang, G., Wang, X., Zhang, Y., Xing, X., Blanton, H., Salem, T. and Jacobs, N. (2020). Joint 2D-3D Breast Cancer Classification. *arXiv:2002.12392*. [online] Available at: https://arxiv.org/abs/2002.12392.

Loizidou, K., Skouroumouni, G., Nikolaou, C. and Pitris, C., 2020. An automated breast micro-calcification detection and classification technique using temporal subtraction of mammograms. *IEEE Access*, *8*, pp.52785-52795.

Logullo, A., Prigenzi, K., Nimir, C., Franco, A. and Campos, M. (2022). Breast microcalcifications: Past, present and future (Review). *Molecular and Clinical Oncology*, 16(4). doi:https://doi.org/10.3892/mco.2022.2514.

Lowd, D. and Domingos, P. (2005). Naive Bayes models for probability estimation. *Proceedings of the 22nd international conference on Machine learning - ICML '05*. doi:https://doi.org/10.1145/1102351.1102418.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J. and Liu, H., 2017. Feature selection: A data perspective. *ACM computing surveys (CSUR)*, *50*(6), pp.1-45.

Li, M., He, L., Lei, C. and Gong, Y. (2021). Fine-grained image classification model based on improved SqueezeNet. *2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*. doi:https://doi.org/10.1109/iaeac50856.2021.9390687.

Lu, Y., Niu, K., Peng, X., Zeng, J. and Pei, S. (2022). Multi-modal Intermediate Fusion Model for diagnosis prediction. In *2022 the 6th International Conference on Innovation in Artificial Intelligence (ICIAI)* (pp. 38-43).

Lu, Z., Lily Rui Liang, Song, G. and Wang, S. (2010). Polychotomous kernel Fisher discriminant via top–down induction of binary tree. *Computers & Mathematics with Applications*, 60(3), pp.511–519. doi:https://doi.org/10.1016/j.camwa.2010.04.048.

Lutfi, M., Hasan Syaiful Rizal, Mochammad Hasyim, Muhammad Faishol Amrulloh and Zulfatun Nikmatus Saadah (2022). Feature Extraction and Naïve Bayes Algorithm for Defect Classification of Manalagi Apples. *Journal of Physics: Conference Series*, 2394(1), pp.012014–012014. doi:https://doi.org/10.1088/1742-6596/2394/1/012014.

Maciej Pawłowski, Wróblewska, A. and Sylwia Sysko–Romańczuk (2023). Effective Techniques for Multimodal Data Fusion: A Comparative Analysis. *Sensors*, 23(5), pp.2381–2381. doi:https://doi.org/10.3390/s23052381.

Magni, V., Cozzi, A., Schiaffino, S., Colarieti, A. and Sardanelli, F. (2023). Artificial intelligence for digital breast tomosynthesis: Impact on diagnostic performance, reading times, and workload in the era of personalized screening. *European Journal of Radiology*, 158, p.110631. doi:https://doi.org/10.1016/j.ejrad.2022.110631.

Mayo Clinic. (n.d.). *Breast MRI*. [online] Available at: https://www.mayoclinic.org/tests-procedures/breast-mri/multimedia/breast-mri/img-20007363 [Accessed 26 Nov. 2023].

Mendel, K., Li, H., Sheth, D. and Giger, M. (2019). Transfer Learning From Convolutional Neural Networks for Computer-Aided Diagnosis: A Comparison of Digital Breast Tomosynthesis and Full-Field Digital Mammography. *Academic Radiology*, 26(6), pp.735–743. doi:10.1016/j.acra.2018.06.019.

Mendes, J., Matela, N. and Garcia, N.C. (2023). Avoiding Tissue Overlap in 2D Images: Single-Slice DBT Classification Using Convolutional Neural Networks. *Tomography*, 9(1), pp.398–412. doi:https://doi.org/10.3390/tomography9010032.

Mindrila, D. and Balentyne, P. (2013). The Chi Square Test. *The Basic Practice of Statistics. 6th ed. New York: WH Freeman*.

Mmed, M. (2023). *Augmenting Medical Imaging: A Comprehensive Catalogue of 65 Techniques for Enhanced Data Analysis*. [online] Available at: https://arxiv.org/pdf/2303.01178.pdf [Accessed 6 Oct. 2023].

Mohammed, A. and Kora, R. (2023). A Comprehensive Review on Ensemble Deep Learning: Opportunities and Challenges. *Journal of King Saud University - Computer and Information Sciences*, 35(2). doi:https://doi.org/10.1016/j.jksuci.2023.01.014.

Mostafa, S. and Wu, F.-X. (2021). Diagnosis of autism spectrum disorder with convolutional autoencoder and structural MRI images. *Neural Engineering*

Techniques for Autism Spectrum Disorder, pp.23–38. doi:https://doi.org/10.1016/b978-0-12-822822-7.00003-x.

Mota, A.M., Clarkson, M.J., Almeida, P. and Matela, N. (2022). Automatic Classification of Simulated Breast Tomosynthesis Whole Images for the Presence of Microcalcification Clusters Using Deep CNNs. *Journal of Imaging*, 8(9), p.231. doi:10.3390/jimaging8090231.

Mota, A.M., Mendes, J. and Matela, N. (2023). Digital Breast Tomosynthesis: Towards Dose Reduction through Image Quality Improvement. *Journal of Imaging*, 9(6), pp.119–119. doi:https://doi.org/10.3390/jimaging9060119.

Muduli, D., Dash, R. and Majhi, B. (2021). Automated diagnosis of breast cancer using multi-modal datasets: A deep convolution neural network based approach. *Biomedical Signal Processing and Control*, p.102825. doi:https://doi.org/10.1016/j.bspc.2021.102825.

mycdiadmin (2016). *What's the Best Cancer Screening for Women with Dense Breasts?* [online] RAYUS Radiology. Available at: https://rayusradiology.com/blog/whats-the-best-cancer-screening-for-women-with-dense-breasts/.

NAACCR, 2022. Cancer In North America, 2015-2019. Volume Three: Registry-specific Cancer Mortality in the United States and Canada. Edited by Sherman R, Firth R, Kahl A, De P, Green D, Hofer B, Liu L, Hsieh M, Johnson C, Kohler B, Morawski B, Nash S, Qiao B, Weir H. Springfield, IL: North American Association of Central Cancer Registries, Inc. Available at: https://www.naaccr.org/cancer-in-north-america-cina-volumes/.

Nakagawa, S. and Cuthill, I.C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews*, 82(4), pp.591–605. doi:https://doi.org/10.1111/j.1469-185x.2007.00027.x.

Nasiri, H. and Alavi, S.A. (2022). A Novel Framework Based on Deep Learning and ANOVA Feature Selection Method for Diagnosis of COVID-19 Cases from Chest X-

Ray Images. *Computational Intelligence and Neuroscience*, 2022, pp.1–11. doi:https://doi.org/10.1155/2022/4694567.

National Cancer Institute. (2018). *Breast Changes and Conditions*. [online] Available at: https://www.cancer.gov/types/breast/breast-changes.

NCHS, 2022. National Center for Health Statistics, Centers for Disease Control and Prevention. Available at: https://www.cdc.gov/nchs/.

Nimish Ukey, Yang, Z., Li, B., Zhang, G., Hu, Y. and Zhang, W. (2023). Survey on Exact kNN Queries over High-Dimensional Data Space. *Sensors*, 23(2), pp.629–629. doi:https://doi.org/10.3390/s23020629.

Nogay, H., Akinci, T.C. and Yilmaz, M. (2021). Comparative Experimental Investigation and Application of Five Classic Pre-Trained Deep Convolutional Neural Networks via Transfer Learning for Diagnosis of Breast Cancer. *Advances in Science and Technology Research Journal*, 15(3), pp.1–8. doi:https://doi.org/10.12913/22998624/137964.

Omer Fadl Elssied, N., Ibrahim, O. and Hamza Osman, A. (2014). A Novel Feature Selection Based on One-Way ANOVA F-Test for E-Mail Spam Classification. *Research Journal of Applied Sciences, Engineering and Technology*, 7(3), pp.625–638. doi:https://doi.org/10.19026/rjaset.7.299.

Paepke, S., Metz, S., Brea Salvago, A. and Ohlinger, R. (2018). Benign Breast Tumours - Diagnosis and Management. *Breast Care*, 13(6), pp.403–412. doi:https://doi.org/10.1159/000495919.

Park, J., Chłędowski, J., Jastrzębski, S., Witowski, J., Xu, Y., Du, L., Gaddam, S., Kim, E., Lewin, A., Parikh, U., Plaunova, A., Chen, S., Millet, A., Park, J., Pysarenko, K., Patel, S., Goldberg, J., Wegener, M., Moy, L. and Heacock, L. (2022). 3D-GMIC: an efficient deep neural network to find small objects in large 3D images. *arXiv:2210.08645*. [online] Available at: https://arxiv.org/abs/2210.08645v1 [Accessed 24 Dec. 2022].

Park, J., Shoshan, Y., Martí, R., Gómez, P., Ratner, V.A., Khapun, D., Zlotnick, A., Barkan, E., Gilboa-Solomon, F., Jakub Chłędowski, Witowski, J., Millet, A., Kim, E.,

Lewin, A.A., Pysarenko, K., Chen, S., Goldberg, J.E., Patel, S., Plaunova, A. and Wegener, M. (2021). Lessons from the first DBTex Challenge. *Nature Machine Intelligence*, 3(8), pp.735–736. doi:https://doi.org/10.1038/s42256-021-00378-z.

Pathan, M.S., Nag, A., Pathan, M.M. and Dev, S., 2022. Analyzing the impact of feature selection on the accuracy of heart disease prediction. *Healthcare Analytics*, *2*, p.100060.

Pawara, P., Okafor, E., Surinta, O., Schomaker, L. and Wiering, M. (2017). Comparing Local Descriptors and Bags of Visual Words to Deep Convolutional Neural Networks for Plant Recognition. *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods*. doi:https://doi.org/10.5220/0006196204790486.

Perez, L., Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. arXiv preprint arXiv:1712.04621.

Peterson, K.D., 2018. Resting heart rate variability can predict track and field sprint performance. *OA Journal-Sports*, *1*(1).

Petrick, N., Sahiner, B., Armato, S.G., Bert, A., Correale, L., Delsanto, S., Freedman, M.T., Fryd, D., Gur, D., Hadjiiski, L., Huo, Z., Jiang, Y., Morra, L., Paquerault, S., Raykar, V., Samuelson, F., Summers, R.M., Tourassi, G., Yoshida, H. and Zheng, B. (2013). Evaluation of computer-aided detection and diagnosis systems). *Medical Physics*, 40(8), p.087001. doi:https://doi.org/10.1118/1.4816310.

Pothos, V., Kastaniotis, D., Theodorakopoulos, I. and Fragoulis, N. (2016). *A fast, embedded implementation of a Convolutional Neural Network for Image Recognition*. Technical Report, Aug. 2016,[Online]. Available: https://www. researchgate. net/publication/306003694_A_fast_embedded_implementation_of_a_Convolutional_Neural_Network_for_Image_Recognition.

Radhakrishna, S., Agarwal, S., Parikh, P., Kaur, K., Panwar, S., Sharma, S., Dey, A., Saxena, K., Chandra, M. and Sud, S. (2018). Role of magnetic resonance imaging in breast cancer management. South Asian Journal of Cancer, 7(2), p.69. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5909298/.

Radwan Qasrawi, Maha Hoteit, Reema Tayyem, Khlood Bookari, Haleama Al Sabbah, Kamel, I., Dashti, S., Sabika Allehdan, Hiba Bawadi, Waly, M.I., Ibrahim, M.O., Charlotte De Backer, Teunissen, L., Kathleen Van Royen, Cuykx, I., Decorte, P., Gaëlle Ouvrein, Karolien Poels, Vandebosch, H. and Katrien Maldoy (2023). Machine learning techniques for the identification of risk factors associated with food insecurity among adults in Arab countries during the COVID-19 pandemic. *BMC Public Health*, 23(1). doi:https://doi.org/10.1186/s12889-023-16694-5.

Ramzan, F., Khan, M.U.G., Rehmat, A., Iqbal, S., Saba, T., Rehman, A. and Mehmood, Z. (2019). A Deep Learning Approach for Automated Diagnosis and Multi-Class Classification of Alzheimer's Disease Stages Using Resting-State fMRI and Residual Neural Networks. *Journal of Medical Systems*, 44(2). doi:https://doi.org/10.1007/s10916-019-1475-2.

Rashidi, H.H., Albahra, S., Robertson, S., Tran, N.K. and Hu, B. (2023). Common statistical concepts in the supervised Machine Learning arena. *Frontiers in Oncology*, 13. doi:https://doi.org/10.3389/fonc.2023.1130229.

Ricciardi, R., Mettivier, G., Staffa, M., Sarno, A., Acampora, G., Minelli, S., Santoro, A., Antignani, E., Orientale, A., Pilotti, I.A.M., Santangelo, V., D'Andria, P. and Russo, P. (2021). A deep learning classifier for digital breast tomosynthesis. *Physica Medica*, 83, pp.184–193. doi:https://doi.org/10.1016/j.ejmp.2021.03.021.

Riffenburgh, R.H. (2006). Statistical Testing, Risks, and Odds in Medical Decisions. *Elsevier eBooks*, pp.93–114. doi: https://doi.org/10.1016/b978-012088770-5/50045-9.

Riggs, M. (2017). Mammograms, are they really preventative? | 5280 Functional Medicine. [online] Available at: https://5280functionalmed.com/are-mammograms-really-preventative/

Rodriguez-Ruiz, A., Teuwen, J., Vreemann, S., Bouwman, R.W., van Engen, R.E., Karssemeijer, N., Mann, R.M., Gubern-Merida, A. and Sechopoulos, I. (2017). New reconstruction algorithm for digital breast tomosynthesis: better image quality for humans and computers. *Acta Radiologica*, 59(9), pp.1051–1059. doi:10.1177/0284185117748487.

Rosenqvist, S., Brännmark, J. and Dustler, M. (2024). Digital breast tomosynthesis in breast cancer screening: an ethical perspective. *Insights into Imaging*, 15(1). doi:https://doi.org/10.1186/s13244-024-01790-w.

R V, Aswiga., R, Aishwarya. and A P, Shanthi. (2021). Augmenting Transfer Learning with Feature Extraction Techniques for Limited Breast Imaging Datasets. *Journal of Digital Imaging*. doi:10.1007/s10278-021-00456-z.

S. Arun Kumar and S. Sasikala (2023). Review on Deep Learning-Based CAD Systems for Breast Cancer Diagnosis. *Technology in Cancer Research & Treatment*, 22, p.153303382311779-153303382311779. doi:https://doi.org/10.1177/15330338231177977.

Saffari, N., Rashwan, H.A., Abdel-Nasser, M., Kumar Singh, V., Arenas, M., Mangina, E., Herrera, B. and Puig, D. (2020). Fully Automated Breast Density Segmentation and Classification Using Deep Learning. *Diagnostics*, 10(11), p.988. doi:https://doi.org/10.3390/diagnostics10110988.

Saifudin, S.A., Sulaiman, S.N., Karim, N.K.A., Osman, M.K., Isa, I.S. and Harron, N.A. (2022). A comparative study of unsharp masking filters for enhancement of digital breast tomosynthesis images. In *2022 IEEE 12th International Conference on Control System, Computing and Engineering (ICCSCE)* (pp. 147-152).

Sakai, A., Onishi, Y., Matsui, M., Adachi, H., Teramoto, A., Saito, K. and Fujita, H. (2019). A method for the automated classification of benign and malignant masses on digital breast tomosynthesis images using machine learning and radiomic features. *Radiological Physics and Technology*, 13(1), pp.27–36. doi:https://doi.org/10.1007/s12194-019-00543-5.

Samala, R.K., Chan, H.-P., Hadjiiski, L.M., Helvie, M.A., Richter, C. and Cha, K. (2018). Evolutionary pruning of transfer learned deep convolutional neural network for breast cancer diagnosis in digital breast tomosynthesis. *Physics in Medicine & Biology*, 63(9), p.095005. doi:10.1088/1361-6560/aabb5b.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L.-C. (2018). *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. [online] arXiv.org. Available at: https://arxiv.org/abs/1801.04381.

Sarker, I.H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, 2(3), pp.1–21. doi:https://doi.org/10.1007/s42979-021-00592-x.

Sarwinda, D., Paradisa, R.H., Bustamam, A. and Anggia, P. (2021). Deep Learning in Image Classification using Residual Network (ResNet) Variants for Detection of Colorectal Cancer. *Procedia Computer Science*, 179, pp.423–431. doi:10.1016/j.procs.2021.01.025.

Sharma, D., Graff, C.G., Badal, A., Zeng, R., Sawant, P., Sengupta, A., Dahal, E. and Badano, A., 2019. In silico imaging tools from the VICTRE clinical trial. *Medical physics*, *46*(9), pp.3924-3928.

Shi, X. and Manduchi, R. (2003). A study on Bayes feature fusion for image classification. In *2003 Conference on Computer Vision and Pattern Recognition Workshop* (Vol. 8, pp. 95-95).

Shoshan, Y., Zlotnick, A., Ratner, V., Khapun, D., Barkan, E. and Gilboa-Solomon, F. (2021). Beyond Non-maximum Suppression - Detecting Lesions in Digital Breast Tomosynthesis Volumes. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp.772–781. doi:10.1007/978-3-030-87240-3_74.

Simonyan, K. and Zisserman, A. (2015). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. [online] arXiv.org. Available at: https://arxiv.org/abs/1409.1556.

Singh, P., Mukundan, R. and De Ryke, R. (2019). Feature Enhancement in Medical Ultrasound Videos Using Contrast-Limited Adaptive Histogram Equalization. *Journal of Digital Imaging*, 33(1), pp.273–285. doi:https://doi.org/10.1007/s10278-019-00211-5.

Singh, S., Matthews, T.P., Shah, M., Mombourquette, B., Tsue, T., Long, A., Almohsen, R., Pedemonte, S. and Su, J. (2020). Adaptation of a deep learning

malignancy model from full-field digital mammography to digital breast tomosynthesis. *arXiv:2001.08381.* [online] Available at: https://arxiv.org/abs/2001.08381v1.

Su, T., Deng, X., Yang, J., Wang, Z., Fang, S., Zheng, H., Liang, D. and Ge, Y. (2021). DIR-DBTnet: Deep iterative reconstruction network for three-dimensional digital breast tomosynthesis imaging. *Medical Physics*, 48(5), pp.2289–2300. doi:https://doi.org/10.1002/mp.14779.

Sulaiman, S.N., Normazli, M.H., Harron, N.A., Karim, N.K.A., Ahmad, K.A. and Soh, Z.H.C. (2022). A Convolutional Neural Network Model for Image Enhancement of Extremely Dense Breast Tissue in Digital Breast Tomosynthesis Images. In *2022 IEEE 12th International Conference on Control System, Computing and Engineering (ICCSCE)* (pp. 153-157).

Sumaiya Thaseen, I. and Aswani Kumar, C. (2017). Intrusion detection model using fusion of chi-square feature selection and multi class SVM. *Journal of King Saud University - Computer and Information Sciences*, 29(4), pp.462–472. doi:https://doi.org/10.1016/j.jksuci.2015.12.004.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2014). *Going Deeper with Convolutions*. [online] arXiv.org. Available at: https://arxiv.org/abs/1409.4842.

Tang, J., Alelyani, S. and Liu, H. (2014). Feature selection for classification: A review. *Data classification: Algorithms and applications*, p.37.

Tardy, M. and Mateus, D. (2021). Trainable Summarization to Improve Breast Tomosynthesis Classification. *Lecture Notes in Computer Science*, pp.140–149. doi:https://doi.org/10.1007/978-3-030-87234-2_14.

Themes, U.F.O. (2016). Digital Mammography and Digital Breast Tomosynthesis. [online] Oncohema Key. Available at: https://oncohemakey.com/digital-mammography-and-digital-breast-tomosynthesis/

The American Cancer Society. (2019). About Breast Cancer. [online] Available at: https://www.cancer.org/content/dam/CRC/PDF/Public/8577.00.pdf.

Tsutomu Gomi, Yukie Kijima, Kobayashi, T. and Yukio Koibuchi (2022). Evaluation of a Generative Adversarial Network to Improve Image Quality and Reduce Radiation-Dose during Digital Breast Tomosynthesis. *Diagnostics*, 12(2), pp.495–495. doi:https://doi.org/10.3390/diagnostics12020495.

Van den Elzen, S. and Van Wijk, J.J. (2011). BaobabView: Interactive construction and analysis of decision trees. *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*. doi:https://doi.org/10.1109/vast.2011.6102453.

Van Gestel, T., Suykens, J.A., Baesens, B., Viaene, S., Vanthienen, J., Dedene, G., De Moor, B. and Vandewalle, J. (2004). Benchmarking least squares support vector machine classifiers. *Machine learning*, *54*, pp.5-32.

Vaswani, A., Brain, G., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł. and Polosukhin, I. (2017). *Attention Is All You Need*. [online] Available at: https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vo, T., Tran, D., Ma, W. and Nguyen, K. (2013). Improved HOG Descriptors in Image Classification with CP Decomposition. *Lecture Notes in Computer Science*, pp.384–391. doi:https://doi.org/10.1007/978-3-642-42051-1_48.

Wang, L. (2019). Research and Implementation of Machine Learning Classifier Based on KNN. *IOP Conference Series: Materials Science and Engineering*, 677, p.052038. doi:https://doi.org/10.1088/1757-899x/677/5/052038.

Wang, L., He, Q., Wang, X., Song, T., Li, X., Zhang, S., Qin, G., Chen, W., Zhou, L. and Zhen, X. (2021). Multi-criterion decision making-based multi-channel hierarchical fusion of digital breast tomosynthesis and digital mammography for breast mass discrimination. *Knowledge Based Systems*, 228, pp.107303–107303. doi:https://doi.org/10.1016/j.knosys.2021.107303.

Williams, L.H. and Drew, T. (2019). What do we know about volumetric medical image interpretation?: a review of the basic science and medical image perception literatures. *Cognitive Research: Principles and Implications*, 4(1). doi:https://doi.org/10.1186/s41235-019-0171-6.

WebMD. (2022). *Are My Breasts Normal or Should I Call the Doctor?* [online] Available at: https://www.webmd.com/women/normal-vs-abnormal-breasts.

Yan, S., Jing, L. and Wang, H. (2021). A New Individual Tree Species Recognition Method Based on a Convolutional Neural Network and High-Spatial Resolution Remote Sensing Imagery. *Remote Sensing*, 13(3), p.479. doi:10.3390/rs13030479.

Yao, W., Moumtzidou, A., Dumitru, C.O., Andreadis, S., Gialampoukidis, I., Vrochidis, S., Datcu, M. and Kompatsiaris, I. (2021). Early and late fusion of multiple modalities in sentinel imagery and social media retrieval. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VII* (pp. 591-606). Springer International Publishing.

Yasaka, K., Akai, H., Kunimatsu, A., Kiryu, S. and Abe, O. (2018). Deep learning with convolutional neural network in radiology. *Japanese Journal of Radiology*, 36(4), pp.257–272. doi:https://doi.org/10.1007/s11604-018-0726-3.

Yeasmin, M.N., Al Amin, M., Joti, T.J., Aung, Z. and Azim, M.A., 2024. Advances of AI in image-based computer-aided diagnosis: A review. *Array*, p.100357.

Zafar, A., Aamir, M., Mohd Nawi, N., Arshad, A., Riaz, S., Alruban, A., Dutta, A.K. and Almotairi, S. (2022). A Comparison of Pooling Methods for Convolutional Neural Networks. *Applied Sciences*, [online] 12(17), p.8643. doi:https://doi.org/10.3390/app12178643.

Zakir Ullah, M., Zheng, Y., Song, J., Aslam, S., Xu, C., Kiazolu, G.D. and Wang, L. (2021). An Attention-Based Convolutional Neural Network for Acute Lymphoblastic Leukemia Classification. *Applied Sciences*, 11(22), p.10662. doi:https://doi.org/10.3390/app112210662.

Zhang, L., Bian, Y., Jiang, P. and Zhang, F. (2023). A Transfer Residual Neural Network Based on ResNet-50 for Detection of Steel Surface Defects. *Applied Sciences*, 13(9), p.5260. doi:https://doi.org/10.3390/app13095260.

Zhang, X., Zhang, Y., Han, E.Y., Jacobs, N., Han, Q., Wang, X. and Liu, J. (2018). Classification of Whole Mammogram and Tomosynthesis Images Using Deep

Convolutional Neural Networks. *IEEE Transactions on NanoBioscience*, 17(3), pp.237–242. doi:10.1109/tnb.2018.2845103.

Zhang, Y., Wang, X., Blanton, H., Liang, G., Xing, X. and Jacobs, N. (2019). 2D Convolutional Neural Networks for 3D Digital Breast Tomosynthesis Classification. *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. doi:10.1109/bibm47256.2019.8983097.

Zheng, C. and Mo, T. (2020). End-to-End Breast Mass Classification on Digital Breast Tomosynthesis.

Appendix A List of Publications

The research presented in this thesis has resulted in the following scholarly publications:

1. El-Shazli, A.M.A., Youssef, S.M. and Soliman, A.H. (2022). Intelligent Computer-Aided Model for Efficient Diagnosis of Digital Breast Tomosynthesis 3D Imaging Using Deep Learning. *Applied Sciences*, 12(11), p.5736. doi:https://doi.org/10.3390/app12115736. (El-Shazli, Youssef and Soliman, 2022)

2. El-Shazli, A.M.A., Youssef, S.M., Soliman, A.H. and Chibelushi C. (2024). An Enhanced Framework Employing Feature Fusion for Effective Classification of Digital Breast Tomosynthesis Scans. *2024 International Conference on Machine Intelligence and Smart Innovation (ICMISI)*, Alexandria, Egypt, 2024, pp. 1-7, doi: 10.1109/ICMISI61517.2024.10580778.

3. El-Shazli, A.M.A., Youssef, S.M., Soliman, A.H. et al. MSAE-DL: enhancing breast cancer classification through hybrid self-attention integration, feature fusion, and ensemble classification in digital breast tomosynthesis. Neural Comput & Applic (2025). https://doi.org/10.1007/s00521-025-11192-8