



Faculty of Computer Science and Information Technology

**A Comparative Study Using Deep Learning Model for BI-RADS
Classification of Mammogram**

Ralf Ashley anak Razali

**Bachelor of Computer Science with Honour
(Computational Science)**

2025

ACKNOWLEDGEMENT

I wish to express my sincere gratitude to my faculty, Faculty of Computer Science and Information Technology (FCSIT), from Universiti Malaysia Sarawak (UNIMAS) for providing the necessary resources, including access to Wi-Fi and essential facilities, which were instrumental in the successful completion of my final year project. The support from the institution has been invaluable in enabling me to carry out my research effectively. I wish to express my sincere gratitude to Dr. Wang Yin Chai, our final-year project coordinator, for offering clear advice and support throughout our endeavour. I am profoundly appreciative of my thesis examiner, Dr. Narayanan Kulathuramaiyer, for dedicating time to evaluate my work and provide helpful input that has greatly enhanced the quality of my thesis.

I express my gratitude to Kaggle and the King Abdulaziz University Mammogram Dataset team for supplying the vital mammogram imaging datasets that were indispensable to my research. Their contributions have been essential in progressing this project. I am deeply appreciative of my friends and family, whose emotional and physical support has consistently provided strength and motivation. Their faith in me and support along this journey have been vital.

ABSTRACT

Breast cancer is one of the major causes of death among women all over the world, and its prognosis largely depends on early detection. Among various techniques used for screening, mammography is a popular one, but its manual interpretation generally leads to variability and diagnostic errors. This study explores the potential of deep learning models in enhancing the accuracy and efficiency of BI-RADS classification for mammograms. For that purpose, three different deep learning architectures are implemented using a publicly available dataset in Kaggle-ResNet, DenseNet, and VGG-and assessed. This preprocessing made data compatible for underrepresented BI-RADS categories to help in class balancing by cleaning, augmenting, and resizing. Accordingly, the models went through hard tests using different metrics such as accuracy, sensitivity, specificity, AUC-ROC. This opens a great opportunity to decrease diagnostic variability, support radiologists in clinical workflows, and thus contribute to the early detection of breast cancer, so important for good patient outcomes. The deployment of the best-performing model on a cloud-based demo platform underlines the practical applicability of this research, allowing stakeholders to interact with the system and assess its diagnostic capability in real-time. These findings demonstrate the transformative potential of deep learning in medical imaging and offer a stable solution for improving the diagnosis of breast cancer, thereby supporting the use of AI-powered tools in healthcare.

Keywords: Breast cancer, mammography, BI-RADS classification, deep learning, DenseNet

ABSTRAK

Kanser payudara merupakan antara punca utama kematian dalam kalangan wanita di seluruh dunia, dan prognosisnya sangat bergantung pada pengesanan awal. Dalam pelbagai teknik saringan yang digunakan, mamografi merupakan antara teknik yang popular, namun tafsiran manualnya sering menyebabkan tiada konsistensi dan kesilapan diagnostik. Kajian ini meneroka potensi model pembelajaran mendalam untuk meningkatkan ketepatan dan kecekapan klasifikasi BI-RADS untuk mamogram. Oleh itu, tiga pembelajaran mendalam termaju telah dilakukan menggunakan data set awam yang tersedia di Kaggle untuk ResNet, DenseNet, dan Inception dinilai. Prapemprosesan data dilakukan untuk memastikan keserasian data bagi kategori BI-RADS yang kurang diwakili melalui pembersihan, pengayaan, dan pengubahan saiz data untuk mengimbangi kelas. Model-model ini kemudiannya diuji dengan teliti menggunakan pelbagai metrik seperti ketepatan, sensitiviti, kekhususan, dan AUC-ROC. Penemuan ini membuka peluang besar untuk mengurangkan ketidakkonsistenan diagnostik, menyokong pakar radiologi dalam aliran kerja klinikal, dan seterusnya menyumbang kepada pengesanan awal kanser payudara, yang penting untuk hasil rawatan pesakit yang berkesan. Penyebaran model dengan prestasi terbaik melalui platform demo berasaskan awan menyerlahkan kebolegunaan penyelidikan ini secara praktikal, membolehkan pihak berkepentingan berinteraksi dengan sistem dan menilai keupayaan diagnostiknya dalam masa nyata. Penemuan ini menunjukkan potensi transformasi pembelajaran mendalam dalam pengimejan perubatan dan menawarkan penyelesaian yang stabil untuk memperbaiki diagnosis kanser payudara, sekali gus menyokong penggunaan alat berkuasa AI dalam penjagaan kesihatan.

Kata kunci: Kanser payudara, mamografi, klasifikasi BI-RADS, pembelajaran mendalam, DenseNet

Table of Contents

Chapter 1 : Introduction.....	1
1.1 Introduction	1
1.2 Problem statement	2
1.3 Objective.....	3
1.4 Methodology.....	4
1.4.1 Business Understanding	4
1.4.2 Data Understanding.....	5
1.4.3 Data Preparation.....	5
1.4.4 Modelling	5
1.4.5 Evaluation.....	6
1.4.6 Deployment.....	6
1.5 Scope	6
1.6 Significance of Project.....	7
1.7 Expected Outcome	7
1.8 Project Schedule.....	9
1.9 Thesis Outline	11
1.9.1 Chapter 1: Introduction	11
1.9.2 Chapter 2: Literature Review	11
1.9.3 Chapter 3: Methodology.....	11
1.9.4 Chapter 4: Experiment and Analysis	11

1.9.5 Chapter 5: Results.....	12
1.9.6 Chapter 6: Summary	12
1.10 Summary	12
Chapter 2 : Literature Review	14
2.1 Introduction	14
2.2 Background knowledge in Domain	14
2.3 Related Works.....	17
2.3.1 Dataset used	17
2.3.2 Preprocessing Techniques.....	18
2.3.3 Machine/Deep Learning Algorithms	19
2.3.4 Parameterization of Algorithms	20
2.3.5 Results	20
2.4 Table of Comparison	22
2.5 Deep Learning Algorithms.....	27
2.5.1 Residual Network (ResNet).....	28
2.5.2 Dense Convolutional Network (DenseNet)	30
2.5.3 Inception	31
2.6 Evaluation Metrics	32
2.6.1 Accuracy.....	32
2.6.2 Area Under the Receiver Operating Characteristic Curve (AUC-ROC)	33
2.6.3 Sensitivity (Recall)	34

2.6.4 Specificity	35
2.7 Tools	36
2.8 Summary	37
Chapter 3 : Methodology	38
3.1 Introduction	38
3.2 Business Understanding	38
3.3 Domain Understanding	39
3.4 Data Preparation	41
3.5 Modelling	42
3.6 Evaluation	43
3.7 Deployment	45
3.8 Summary	46
Chapter 4 : Implementation	47
4.1 introduction	47
4.2 data preparation	47
4.2.1 Class Imbalance and Offline Augmentation	48
4.2.2 Image Resizing Strategy	48
4.2.3 Resizing and Augmentation Functions	49
4.3 Modelling	51
4.3.1 Modelling with DenseNet121	52
4.3.2 Modelling with InceptionV3	54

4.3.3 Modelling with ResNet50	55
4.4 Evaluation.....	57
4.4.1 Accuracy.....	57
4.4.2 F1-Score	58
4.4.3 Recall.....	59
4.4.4 Specificity	59
4.4.5 ROC-AUC.....	60
4.4.6 Confusion Metrix	60
4.4.7 Accuracy and Loss over Epochs	61
4.5 Deployment.....	62
4.5.1 Deployment Platform.....	62
4.5.2 Application Interface Key Components.....	63
4.6 Conclusion	65
Chapter 5 : Result and Discussion	66
5.1 Introduction	66
5.2 Confusion Matrix.....	66
5.3 Classification Metric.....	69
5.4 ROC-AUC Graph.....	72
5.5 Accuracy and Loss over Epochs	75
5.6 Comparative Analysis of Models.....	78
5.7 Conclusion	79

Chapter 6 : Conclusion and Future Work	81
6.1 Introduction	81
6.2 Contribution.....	81
6.3 Limitation	82
6.4 Future works	83
6.5 Conclusion	85
References	86

List of Figure

Figure 1.1 Methodology diagram.....	4
Figure 1.2 Gantt Chart for Project Schedule FYP 1	9
Figure 1.3 Gantt Chart for Project Schedule FYP 2	10
Figure 2.1 Show normal breast (left) and abnormal breast (right) using mammograms.	16
Figure 2.2 Architecture of ResNet 18 that has been customized	29
Figure 3.1 Model of CRISP-DM	38
Figure 4.1 Function to resize and pad images.....	49
Figure 4.2 Function to crop the ROI and Zoom area	50
Figure 4.3 Function to apply offline augmentation	51
Figure 4.4 DenseNet121 based model architecture	52
Figure 4.5 Two phase training with fine tuning for DenseNet121.....	53
Figure 4.6 InceptionV3 based model architecture.....	54
Figure 4.7 Two phase training with fine tuning for InceptionV3.....	54
Figure 4.8 ResNet50 based model architecture	55
Figure 4.9 Two-phase training with fine-tuning for ResNet50.....	56
Figure 4.10 Performance Metric for all three models.....	57
Figure 4.11 ROC-AUC score calculation for multi-class classification	60
Figure 4.12 Confusion matrix visualization using heatmap	60
Figure 4.13 Training and validation accuracy and loss plots	61
Figure 4.14 Loading saved DenseNet model	63
Figure 4.15 Prediction function for class probabilities	64
Figure 4.16 Gradio interface setup for BI-RADS classification demo	64

Figure 5.1 ROC-AUC Graph for DenseNet121 model.....	72
Figure 5.2 ROC-AUC Graph for Inceptionv3 model.....	73
Figure 5.3 ROC-AUC Graph for ResNet50 model.....	74
Figure 5.4 Accuracy and Loss over Epochs for DenseNet121 model.....	75
Figure 5.5 Accuracy and Loss over Epochs for InceptionV3 model.....	76
Figure 5.6 Accuracy and Loss over Epochs for Resnet50 model	77

List of Table

Table 2.1 Comparison of Findings in Related Works	22
Table 2.2 Summary of the tools used in the research paper	36
Table 4.1 Class Distribution before and after augmentation and resizing.....	49
Table 5.1 Confusion Matrix for DenseNet121 model	67
Table 5.2 Confusion Matrix for InceptionV3 model	67
Table 5.3 Confusion Matrix for ResNet50 model	68
Table 5.4 Classification Metric for DenseNet121 models.....	69
Table 5.5 Classification Metric for InceptionV3 model	70
Table 5.6 Classification Metric for ResNet50.....	71
Table 5.7 Details Analysis for all three models	78

List of Equation

Equation 2.1 Equation of Accuracy	32
Equation 2.2 Equation for AUC-ROC.....	34
Equation 2.3 Equation for Sensitivity	35
Equation 2.4 Equation for Specificity	35

Chapter 1: Introduction

1.1 Introduction

Breast cancer is one of the leading causes of death among women globally. The initial report occurred during ancient Egyptian times, almost 3,500 years ago, around 1500 B.C. (Bhushan et al., 2021). Today, it is emerging as a primary cause of mortality among women and is the second most prevalent cancer (Bhushan et al., 2021). In 2018, Malaysia reported 7,539 occurrences of breast cancer, with 2,894 fatalities attributed to the disease during the same year (Tan et al., 2021). In the United States, about 281,550 women were diagnosed with breast cancer in 2021, and the American Cancer Society anticipated that 43,600 women died to the disease (Bhushan et al., 2021). Invasive ductal carcinomas (IDCs) comprise up to 80% of invasive breast cancer instances. Pervasive lobular carcinoma is the second most prevalent type of cancer. More than 80% of non-invasive in situ cancers are ductal, while around 10% are lobular (Watkins, 2019). Breast cancers may be categorized as benign or malignant. Malignant tumours are invasive (cancerous), while benign tumours are non-invasive (noncancerous). The invasive phase of a disorder ascertains its benignity or malignancy. Both tumours include additional subgroups necessitating distinct diagnoses (Ramadhani et al., 2021). Consequently, early detection is essential to reduce the incidence of breast cancer. Mammography is one of the most effective tools for early detection.

Mammography generates breast images via low-dose X-rays, assisting radiologists in identifying abnormalities such as tumours, calcifications, and distortions (Wang, 2024). Consequently, the sole methods for identifying microcalcifications and masses as preliminary indicators of breast cancer are via imaging techniques. Although additional imaging modalities exist, Full Field Digital Mammography (FFDM) is the sole method clinically sanctioned for

population-based breast cancer screening (Ramadhani et al., 2021). Numerous deep architectures have been analysed and effectively employed in breast cancer research, producing favourable outcomes. The Breast Imaging Data and Reporting System (BI-RADS) scoring system is essential in identifying the type of breast cancer (Muhammad Mulana Ramadan Purnama et al., 2024). BI-RADS provides guidelines and criteria for physicians to classify breast cancer cases in clinical imaging. The American College of Radiology (ACR) created BI-RADS to standardize the terminology for mammography reporting and result assessment (Muhammad Mulana Ramadan Purnama et al., 2024). Consequently, improving the accuracy and automating the management of medical data at each stage of the therapeutic process can produce beneficial results (Boroumandzadeh & Parvinnia, 2021).

1.2 Problem statement

Breast cancer is a most commonly cancer that happen to a woman around the world. The pathogenesis of breast cancer is complex and not fully comprehended, however specific risk factors are identified. Increasing age and feminine gender are the predominant risk factors. Genetic alterations, particularly the Breast Cancer genes (BRCA) 1 and 2, are responsible for approximately 10% of breast cancer cases (Watkins, 2019). Additional recognized risk factors encompass a history of ductal carcinoma in situ, elevated body mass index (BMI), first childbirth after the age of 30 or nulliparity, early menarche (prior to age 13), familial history of breast or ovarian cancer, late menopause, and the utilization of postmenopausal hormone therapy(Watkins, 2019).The treatment of breast cancer necessitates a multidisciplinary team of experts in medical, surgical, and radiation oncology. As most breast cancers are detected by mammography or physical examination, primary care physicians frequently serve as the initial point of contact for patients receiving a new breast cancer diagnosis. (Trayes & Cokenakes, 2021).

Mammography is a prevalent imaging technique utilized for breast cancer screening. The procedure includes compressing the breast between two plates while an X-ray beam traverses it, recording variations in the absorption of X-rays by different tissues, including fatty, glandular, and malignant tissues. A digital detector acquires the X-rays that traverse, producing a two-dimensional image of the breast tissue for assessment (Wang, 2024). The staging of breast cancer is based on tumour dimensions, lymph node involvement, the existence of metastases, and certain biomarkers ((Trayes & Cokenakes, 2021). An adaptive transfer learning deep convolutional neural network designed for segmenting mammography pictures containing calcifications, intended to facilitate early breast cancer diagnosis and intervention. Different DL models exhibit strengths and weaknesses in medical image classification, particularly in BI-RADS classification for breast cancer diagnosis (Gao et al., 2023). Some models, such as ResNet and DenseNet, are known for their ability to extract deep hierarchical features, while others, like Inception, offer simpler yet effective feature extraction. Additionally, computational efficiency, accuracy, sensitivity, and specificity vary between models, making it crucial to evaluate and compare their performance comprehensively. Researchers performed visual and quantitative analyses of pre-processed and post-processed photos, demonstrating the model's efficacy in detecting and diagnosing microcalcifications. The introduction of a U-Net-based model for the diagnosis of CAD systems in digital mammograms, facilitating temporal lesion tracking (Gao et al., 2023).

1.3 Objective and Research Study

The objective of this project is:

1. To preprocess the mammogram images for BIRADS classification.

2. To compare the performances of deep learning models for BIRADS classification.
3. To determine the best deep learning model for BIRADS classification.

The research study of this project is:

1. How can deep learning models improve the accuracy of BI-RADS classification for mammograms?
2. What preprocessing techniques are most effective in optimizing mammogram images for classification?
3. How can automated BI-RADS classification assist radiologists in reducing diagnostic variability and errors?

1.4 Methodology

This project follows a data science life cycle encompassing multiple critical phases, ranging from data collecting to model implementation or demonstration as shown in the Figure 1 below:



Figure 1.1 Methodology diagram

1.4.1 Business Understanding

The BI-RADS classification system is essential for standardising mammographic assessments and guiding clinical decisions regarding breast cancer diagnosis. However, manual interpretation of mammograms can introduce variability and errors, impacting diagnostic accuracy. This study focuses on leveraging deep learning models to enhance the precision and

efficiency of BI-RADS classification. By comparing the performance of deep learning, the research aims to identify the most suitable model that can contribute to improved clinical workflows and patient outcomes.

1.4.2 Data Understanding

The dataset used in this study is sourced from Kaggle and comprises mammographic images annotated with BI-RADS categories. A thorough analysis of the dataset is conducted to understand its structure, distribution, and potential biases. The dataset is examined for variations in image quality, class imbalance, and other characteristics that may influence model performance.

1.4.3 Data Preparation

Data preprocessing involves cleaning, resizing, and normalizing mammographic images to ensure consistency and compatibility with the selected CNN architectures. Techniques such as data augmentation are applied to increase the diversity of the training data and mitigate overfitting, while any missing or noisy data are addressed to enhance the reliability of the dataset.

1.4.4 Modelling

Three deep learning architectures are implemented to train models for BI-RADS classification. Each model is optimized through hyperparameter tuning and trained using the preprocessed dataset. The choice of these architectures is motivated by their established performance in medical imaging and classification tasks.

1.4.5 Evaluation

The trained models are evaluated using standard performance metrics, such as accuracy, precision, recall, F1-score, and AUC-ROC. These metrics provide a comprehensive assessment of each model's capability to classify BI-RADS categories effectively. Comparative analysis of the results is conducted to determine the most accurate and reliable model.

1.4.6 Deployment

The selected model is designed for deployment in a clinical setting, where it can assist radiologists in automating BI-RADS classification. Deployment considerations include model scalability, integration into existing diagnostic systems, and user interface design to ensure accessibility and usability in real-world applications.

1.5 Scope

This research will concentrate on the analysis of mammography images categorised under BIRADS levels 1, 3, 4, and 5. These levels denote a spectrum from normal results (BIRADS 1) to a high suspicion of cancer (BIRADS 5), which are essential for automated classification. Multiple image segmentation approaches, including thresholding, region-based algorithms, and clustering, will be employed to delineate areas of interest (ROI) within the mammograms. The study will compare the performance of various deep learning models, assessing them according to critical metrics such as accuracy, sensitivity, and specificity. The study will investigate deep learning-based segmentation techniques for the automated identification and extraction of questionable regions in mammograms.

This project will exclude BIRADS levels 0, 2, and 6. BIRADS 0 signifies inconclusive results necessitating additional imaging, but BIRADS 2 denotes benign findings that do not warrant further assessment, rendering them not as relevant to the present study. BIRADS 6

refers to proven cancer cases, which fall beyond the parameters of this research, as the primary focus is on detection and classification rather than established diagnosis. The research will avoid manual readings by radiologists and is exclusive to breast cancer detection, omitting the examination of other cancer kinds.

1.6 Significance of Project

This project can majorly impact various stakeholders. Enhanced diagnostic precision and early identification of breast cancer by deep learning models can result in superior treatment outcomes and elevated survival rates for patients. Healthcare providers can gain from these models, which aid radiologists in interpreting mammograms with greater consistency and accuracy, hence decreasing the likelihood of missed diagnoses and minimising unnecessary biopsies. Moreover, the project can elevate community knowledge of the significance of frequent mammograms for early detection, hence encouraging increased female participation in screening programs and bolstering public health activities.

1.7 Expected Outcome

One of the primary results of this project will be the identification of the most critical features in mammogram images that significantly lead to the accurate classification of BIRADS. These qualities may encompass elements such as tissue density, texture patterns, and areas of concern, which are essential in assessing whether a mammography reveals normal tissue or potential cancer. Comprehending these properties will improve the interpretability of deep learning models and offer insights into the factors that influence accurate predictions.

The study will offer a thorough comparison of the performance of several deep learning models, including a full assessment of their accuracy, AUC-ROC (Area Under the Receiver Operating Characteristic Curve) sensitivity, and specificity in predicting BIRADS categories.

This comparison will ascertain which model architecture (e.g., CNN, ResNet, or DenseNet) is most efficacious for BIRADS classification, finally choosing the model that provides the maximum prediction performance and reliability. These findings will guide future initiatives to incorporate such models into clinical processes for enhanced breast cancer screening and diagnosis.

1.8 Project Schedule

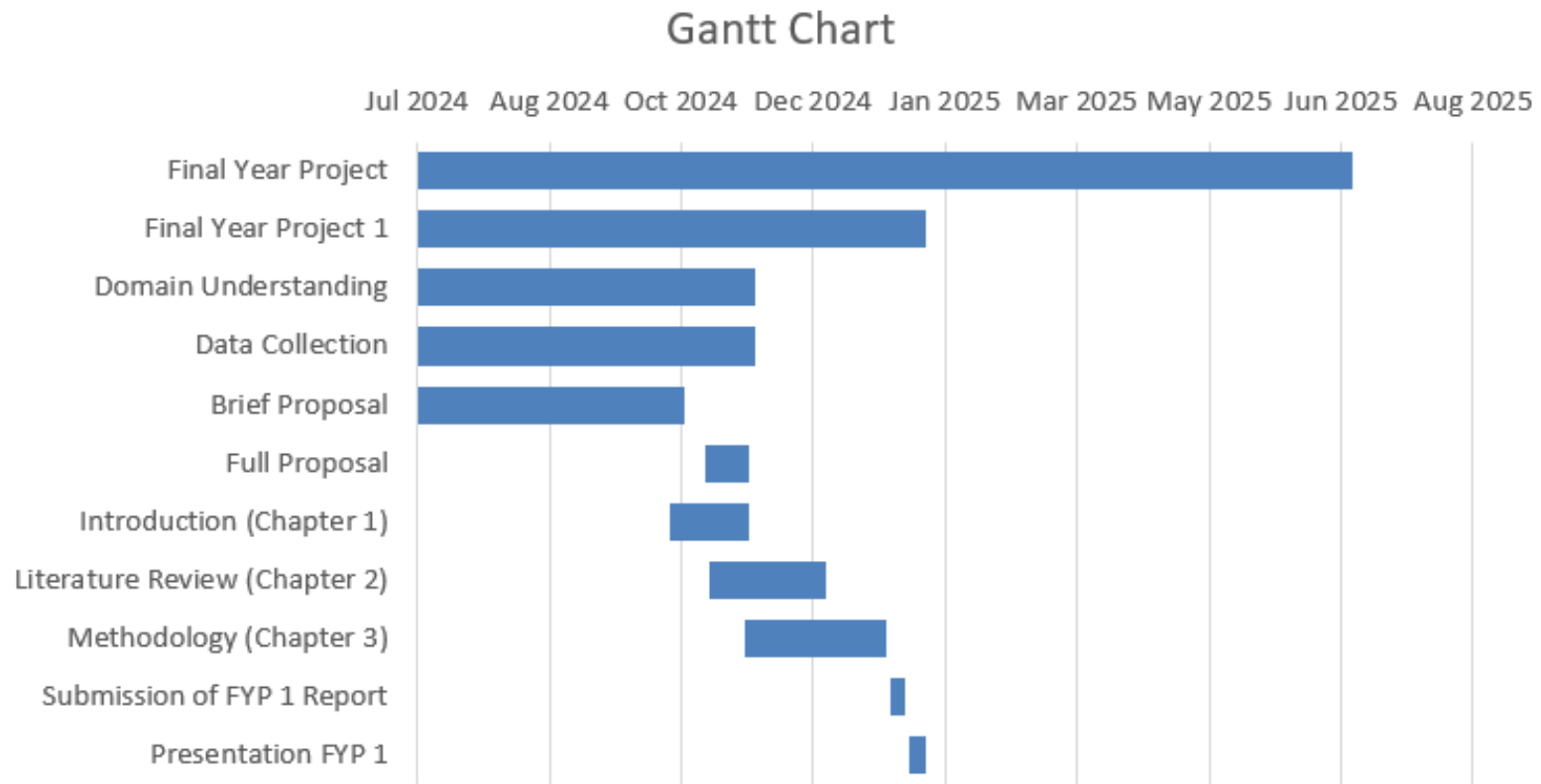


Figure 1.2 Gantt Chart for Project Schedule FYP 1

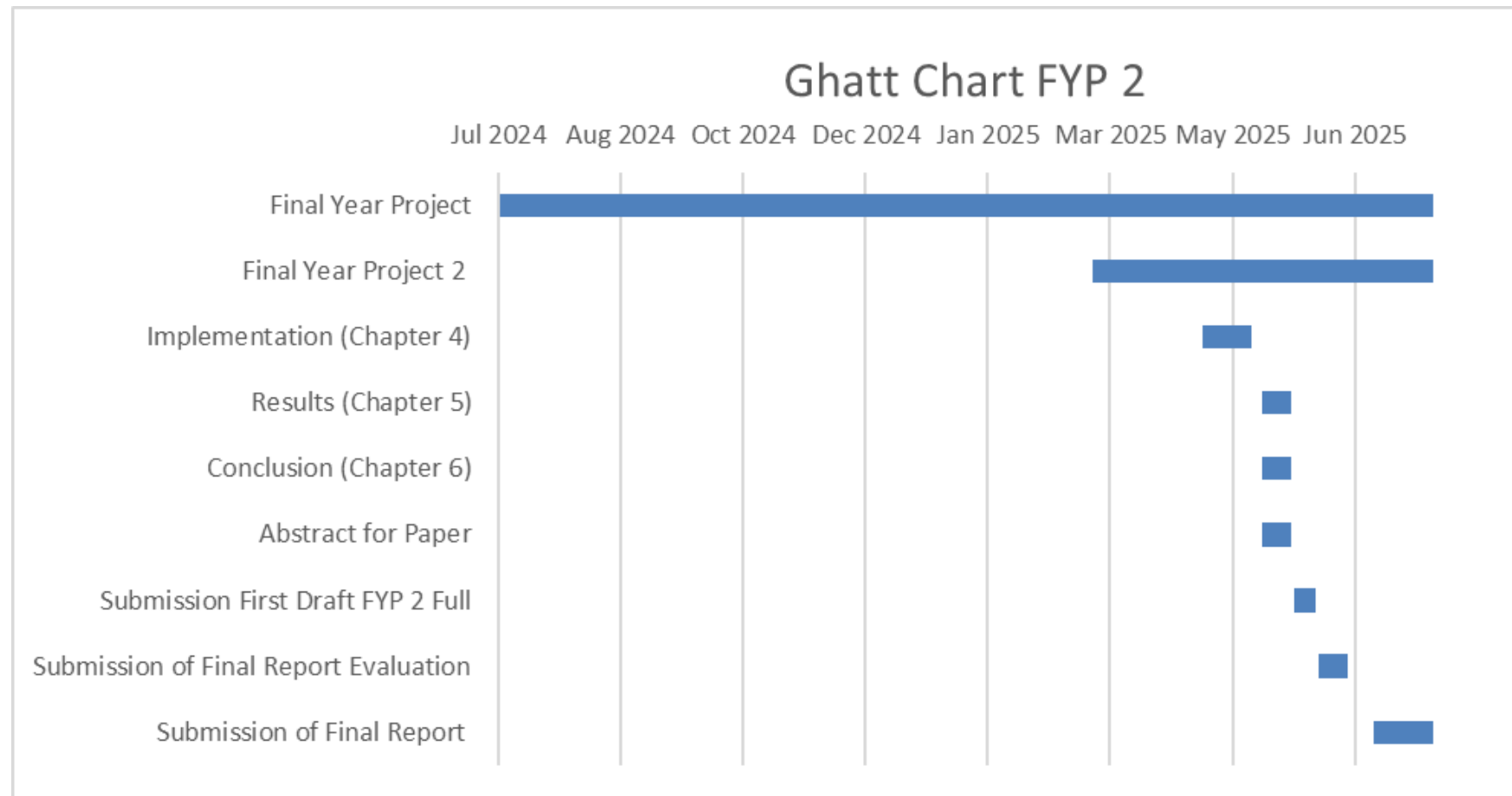


Figure 1.3 Gantt Chart for Project Schedule FYP 2

1.9 Thesis Outline

1.9.1 Chapter 1: Introduction

This chapter presents a comprehensive overview of the project, encompassing the background, problem description, objectives, methods, scope, relevance, and expected results. This establishes the foundations for the research by emphasising the significance of automating BIRADS categorisation of mammograms with deep learning.

1.9.2 Chapter 2: Literature Review

This chapter provides a comprehensive review of previous research concerning BIRADS classification, mammography analysis, and deep learning models. It emphasises the principal advancements, current obstacles, and the justification for employing deep learning for this purpose. This evaluation contextualises the project's methodology and identifies deficiencies in the existing literature.

1.9.3 Chapter 3: Methodology

This chapter explores the methodologies taken to accomplish the project objectives, encompassing data preparation, feature selection, model construction, and evaluation. The chapter discusses the data science life cycle employed in the project and delineates the implementation and testing of various deep learning models.

1.9.4 Chapter 4: Experiment and Analysis

This chapter explores the experiments conducted to evaluate the performance of several deep learning models. This document outlines the setup, execution, and outcomes of the tests,

accompanied by a comprehensive analysis of the models' accuracy, sensitivity, and specificity in BIRADS categorisation.

1.9.5 Chapter 5: Results

This chapter examines the results of the experiments, thereby including the identification of significant predictive factors and the comparative efficacy of several deep learning models. It emphasises the most effective model and offers insights into its efficacy for BIRADS categorisation.

1.9.6 Chapter 6: Summary

The last chapter summarises the project findings, examines the research's overall contributions, and explores prospective projects. It underscores the project's importance in enhancing diagnostic procedures and directing future research in medical image processing through deep learning.

1.10 Summary

This chapter presents breast cancer as the main cause of mortality among women worldwide, examining its historical origins in ancient Egypt and emphasising its current prevalence. The chapter delineates breast cancer classifications, chiefly invasive ductal carcinomas (IDCs) and lobular carcinoma, while underscoring the significance of early identification by mammography. Full Field Digital Mammography (FFDM) is recognised as a principal screening technique that assists in the diagnosis of aberrant calcifications and masses. The BI-RADS rating system of the American College of Radiology is an essential instrument for standardising breast cancer evaluations. The project utilises deep learning architectures to enhance breast cancer detection by training models using mammography datasets, with the goal

of identifying the most effective model for BI-RADS classification to aid radiologists in practical environments.

The chapter outlines the study's aims, which encompass the preprocessing of mammography images and the comparison of deep learning models, including DenseNet, ResNet, and CNN, for accuracy in BI-RADS classification. Data preprocessing strategies, feature selection, and model training methodologies utilising TensorFlow and PyTorch are delineated to guarantee robust model performance. The project's scope concentrates on BI-RADS levels 1, 3, 4, and 5, excluding others due to their minimal significance for early diagnosis. This project seeks to enhance diagnostic accuracy and raise community knowledge regarding the significance of frequent mammograms, ultimately improving early detection and bolstering public health activities for the benefit of patients and healthcare providers.

Chapter 2: Literature Review

2.1 Introduction

Early identification of breast cancer is crucial for enhancing treatment outcomes and survival rates. Mammography is the predominant screening method for the early detection of breast cancer, frequently evaluated through the Breast Imaging Reporting and Data System (BIRADS). The BIRADS classification categorises results by levels of suspicion, assisting radiologists in diagnosis and patient management. Mammogram interpretation is intricate and susceptible to variability stemming from disparities in radiologists' experience, weariness, and subjective assessment. This diversity may lead to inconsistent classifications, affecting the accuracy and reliability of breast cancer diagnosis. Recent breakthroughs in deep learning have yielded interesting techniques for automating mammography classification, potentially enhancing consistency and accuracy in BIRADS categorisation.

This literature review chapter will analyse multiple research projects utilising deep learning models for mammography analysis, emphasising their data sources, preprocessing methods, selected algorithms, and model parameterizations. It will also offer a comparative analysis of performance parameters such as accuracy, sensitivity, and specificity, which are essential for assessing model efficacy. This chapter summarises relevant works to identify current trends, difficulties, and research gaps between 2019 until 2023, therefore contextualising the methods employed in this study.

2.2 Background knowledge in Domain

Breast Cancer is a world known common diagnosed cancer that happen worldwide including Malaysia. In 2020, an estimation of women that died due to breast cancer was 685,000

(Bhushan et al., 2021). Some leading factor that can cause breast cancer is older age, female sex, early menarche, late menopause, family history, dense breast tissue and some other factors (Trayes & Cokenakes, 2021). Consequently, early diagnostic initiatives should be prioritised over opportunistic or organised, population-based screening until the necessary infrastructure and organisational prerequisites for screening are established to accommodate this supplementary activity. Health planners, policymakers, and other stakeholders, including clinicians, educators, community members, and advocates, must comprehend the health system requirements and overall costs associated with breast cancer early detection strategies to formulate effective investments, plans, and policies (Ginsburg et al., 2020). By using BI-RADS, we can classify the breast cancer into different levels to determine which action we should take. To determine the BI-RADS level, we can use mammograms. Mammograms is frequently employed to detect breast cancer. If the screening mammography revealed an anomaly in the breast tissues, a diagnostic mammogram is warranted. It is often advised to conduct a more thorough investigation of the suspected matter. The initial indication of breast cancer is typically a lump in the breast. Breast or underarm discomfort that persists beyond menstruation(Ramadan, 2020). Typically, these lumps can be identified through screening mammography well in advance of the patient's ability to perceive them, even if they are too small to cause any noticeable changes in the patient(Ramadan, 2020). Figure 3 shows the difference between normal breasts and abnormal breasts using mammograms.

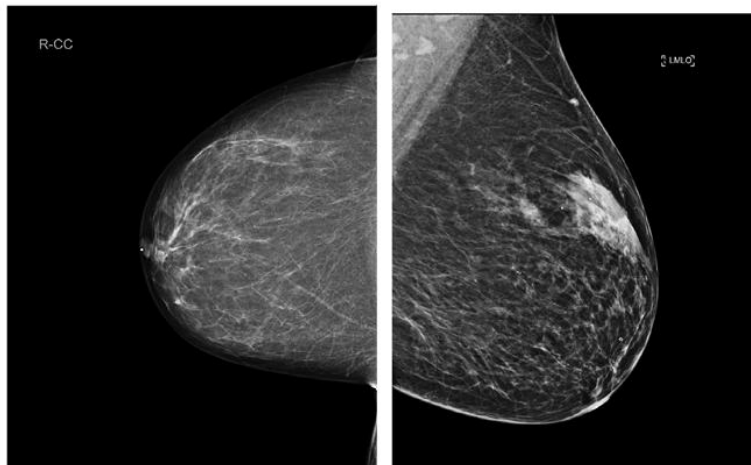


Figure 2.1 Show normal breast (left) and abnormal breast (right) using mammograms.

The use of deep learning in medical imaging has transformed medical image analysis by delivering superior performance in classification, detection, and segmentation tasks. These models can discern intricate patterns from extensive datasets, rendering them exceptionally beneficial for medical applications. Nonetheless, obstacles such as the requirement for extensive annotated datasets, model interpretability, and generalisability across various imaging devices and institutions persist (Pullaiah et al., 2022; Roth et al., 2020). The integration of deep learning with BIRADS classification has demonstrated encouraging outcomes in enhancing the precision and uniformity of breast cancer diagnosis. Utilising the BIRADS-SSDL Network, a semi-supervised deep learning framework that integrates BIRADS characteristics for the diagnosis of breast ultrasound images (Zhang & Gu, 2020). This methodology employs BIRADS-oriented feature maps as input for a stacked convolutional autoencoder, resulting in elevated classification accuracy and generalisability across various datasets (Zhang & Gu, 2020). An additional integration that can be utilised Customised Convolutional Neural Network based on ResNet Employing tailored deep learning architectures to enhance breast cancer categorisation based on BIRADS characteristics. These models have demonstrated enhanced

performance relative to traditional CNNs, especially when employing methods such as batch normalisation and data augmentation (Pullaiah et al., 2022). Visually Interpretable Networks develop deep learning frameworks that offer visual elucidations of diagnostic conclusions grounded around BIRADS criteria. This method improves the interpretability of AI models, rendering them more appropriate for clinical use (Kim et al., 2018).

2.3 Related Works

2.3.1 Dataset used

The search employed diverse datasets, each designed for mammographic image analysis tasks. Notable among these were the Digital Database for Screening Mammography (DDSM), INbreast, and the Mammographic Image Analysis Society (MIAS) dataset. The DDSM, with over 10,000 annotated photos, functioned as an essential resource for the training and validation of deep learning models. It provided varied representations of BI-RADS categories, facilitating models to generalise effectively across lesion kinds. (H. Liu et al., 2021) INbreast, comprising 410 high-resolution pictures and meticulous radiologist annotations, is optimal for jobs necessitating thorough ground truth. The MIAS dataset, comprising 322 greyscale photos, was often utilised for benchmarking owing to its simplicity and accessibility. Custom clinical datasets, including 414 lesions from 384 patients, were utilised in specific investigations to tackle diagnostic problems such as malignancy prediction in BI-RADS 4 microcalcifications (H. Liu et al., 2021).

These datasets frequently necessitated augmentation to mitigate issues of limited size and diversity. Methods such as rotations, flips, and synthetic sample production enhanced the models' robustness. Moreover, datasets were generally partitioned into training, validation, and test subsets to enable impartial assessment. Notwithstanding their utility, these datasets

exhibited issues including inconsistencies in image quality, resolution, and noise levels. Studies utilising clinical datasets have encountered challenges in standardisation owing to discrepancies in imaging techniques. Confronting these problems necessitated careful preprocessing and stringent data selection procedures, as evidenced in several investigations (Daniel López-Cabrera et al., 2020; Pullaiah et al., 2022; Ramadhani et al., 2021).

2.3.2 Preprocessing Techniques

Preprocessing was essential for preparing mammographic pictures for analysis, guaranteeing that the input data was clean, consistent, and diagnostically pertinent. Numerous studies employed manual segmentation by radiologists to define areas of interest (ROIs), such as microcalcifications, and to eliminate extraneous breast tissue. This methodology was especially beneficial for smaller datasets, since a concentrated emphasis on diagnostically pertinent variables could greatly enhance model efficacy (H. Liu et al., 2021). Standardisation methods, including resizing images to consistent size (e.g., 224x224 or 512x512 pixels) and normalising pixel intensity values, guaranteed compatibility across various deep learning systems (Ramadhani et al., 2021). Data augmentation techniques, including as rotation, flipping, cropping, and contrast enhancement, were employed to artificially boost dataset variety and reduce overfitting (Daniel López-Cabrera et al., 2020; Gandomkar et al., 2019).

Alongside these conventional methodologies, certain research utilised sophisticated preprocessing methods to improve image quality and concentrate on diagnostically significant regions (H. Liu et al., 2021). Techniques such as Contrast Limited Adaptive Histogram Equalisation (CLAHE) enhanced the visibility of small lesions, whereas noise reduction strategies mitigated fluctuation in image quality. In cases of imbalanced datasets, synthetic

oversampling techniques such as SMOTE were employed to guarantee equitable representation of under-represented lesion categories. The preprocessing procedures guaranteed that models were trained on high-quality data, improving their capacity to generalise across diverse BI-RADS categories and imaging settings (Sabani et al., 2022; Wang, 2024).

2.3.3 Machine/Deep Learning Algorithms

Convolutional Neural Networks (CNNs) were the predominant deep learning methodology in the investigations, owing to their capacity to extract hierarchical characteristics from mammographic pictures. Architectures such as ResNet, DenseNet, and MobileNetV2 were frequently utilised, optimising computing efficiency and diagnostic precision. In one study, MobileNetV2 was used with clinical data to predict malignancy in BI-RADS 4 lesions, resulting in an AUC of 0.910 (H. Liu et al., 2021). Transfer learning was commonly employed, utilising pre-trained models such as InceptionV3, which were initially trained on ImageNet and subsequently fine-tuned on domain-specific datasets. This method allowed models to utilise established feature extraction abilities and adjust to mammographic imaging tasks, markedly enhancing performance (Daniel López-Cabrera et al., 2020).

Hybrid models that combine imaging characteristics with clinical data, including patient age and menopausal status, significantly improved diagnostic skills. These models frequently surpassed independent imaging models by offering a comprehensive perspective of the patient's condition (Ramadhani et al., 2021). Several studies investigated conventional machine learning algorithms, such as support vector machines (SVMs) and random forests, notably as benchmarks or for datasets where deep learning was impractical due to size limitations. The selection of the method was highly correlated with the characteristics of the dataset, available

processing resources, and the complexity of the diagnostic task (Pullaiah et al., 2022; Wang, 2024).

2.3.4 Parameterization of Algorithms

Hyperparameter adjustment was essential for enhancing model performance and reducing overfitting. Learning rates were generally established between 0.001 and 0.005 to guarantee consistent convergence during training, whereas dropout rates varied from 0.5 to 0.6 to mitigate overfitting, especially in trials with smaller datasets (H. Liu et al., 2021). Methods such as grid search and random search were frequently employed to determine optimal configurations for learning rates, batch sizes, and layer quantities (Sabani et al., 2022). Cross-validation and early stopping techniques were utilised to avert overfitting of models to the training data and to guarantee strong generalisation to novel data (Daniel López-Cabrera et al., 2020; Khanna et al., 2022).

Transfer learning significantly improved parameterisation, especially when utilising pre-trained architectures like MobileNetV2 and InceptionV3. These models preserved generalised characteristics from extensive datasets such as ImageNet, while task-specific layers were optimised for mammographic imaging (Pullaiah et al., 2022). Optimisation techniques like Adam and stochastic gradient descent (SGD) facilitated efficient weight updates, balancing convergence rate with precision. Thorough parameterisation, coupled with effective optimisation procedures, enabled models to attain superior performance across various mammographic datasets and diagnostic tasks (Khanna et al., 2022; Wang, 2024)

2.3.5 Results

The research regularly showed substantial improvements in diagnostic precision and dependability through the application of deep learning models. Performance was assessed using

metrics including accuracy, sensitivity, specificity, and AUC-ROC. A model integrating MobileNetV2 and clinical data attained a sensitivity of 85.3% and specificity of 91.9%, with an AUC of 0.910, surpassing independent imaging models and closely aligning with the performance of professional radiologists (H. Liu et al., 2021). A separate study utilising InceptionV3 with transfer learning achieved an accuracy of 91.3% for malignant/benign classification and an AUC of 93.3%, underscoring the model's efficacy in lesion detection and classification (Daniel López-Cabrera et al., 2020).

Comparative assessments underscored the significance of dataset size and diversity in attaining optimal outcomes. Models trained on extensive datasets like as DDSM and INbreast typically surpassed those trained on smaller datasets like MIAS, highlighting the paramount importance of data quality (Sabani et al., 2022). Moreover, hybrid models that include clinical and imaging data typically outperformed independent methods, underscoring the efficacy of multi-modal research. The results demonstrated the diagnostic capabilities of deep learning and highlighted its significance as a decision-support instrument in clinical practice, especially in high-demand situations (Gao et al., 2023; Pullaiah et al., 2022)

2.4 Table of Comparison

Table 2.1 Comparison of Findings in Related Works

No	Study Name	Description	Deep Learning Model/Algorithm	Evaluation Metrics	Results	Conclusion
1	(H. Liu et al., 2021)	- Used BI-RADS 4 microcalcifications dataset with 414 lesions. - Manual segmentation ensured focus on ROIs. - MobileNetV2 integrated with clinical features. - Achieved AUC of 0.910 with high sensitivity and specificity.	MobileNetV2	AUC, Sensitivity, Specificity	AUC: 0.910, Sensitivity: 85.3%, Specificity: 91.9%	- Integrated clinical and imaging data achieved high diagnostic accuracy, comparable to senior radiologists, emphasizing practical utility in clinical settings.
2	(Ramadhani et al., 2021)	- Utilized DDSM, INbreast, and MIAS datasets. - Standardized image preprocessing (scaling, normalization). - Applied CNN architectures like ResNet and DenseNet.	ResNet50, InceptionV3	Accuracy, Sensitivity, Specificity	Accuracy >90%, Sensitivity >95%	- CNN-based methods like ResNet and DenseNet showed robust performance in BI-RADS classification but highlighted challenges like dataset variability and limited generalizability.

		- Highlighted improvements in BI-RADS classification.				
3	(Sabani et al., 2022)	- Focused on soft tissue opacity classification in BI-RADS categories. - Pre-processed with normalization and augmentation. - Explored ResNet and DenseNet architectures. - Addressed dataset discrepancies with robust preprocessing.	Deep Convolutional Neural Network (CNN)	Accuracy, AUC	Accuracy >90%	- Focused on soft tissue opacity classification and demonstrated high accuracy, emphasizing the importance of robust preprocessing and tailored CNN architectures.
4	(Gandomkar et al., 2019)	- Used DDSM and INbreast datasets for density categorization. - Applied CLAHE for contrast enhancement. - Utilized ResNet-based CNN architecture. - Showcased robust handling of smaller datasets.	Deep Neural Networks (DNN)	Accuracy, Sensitivity, Specificity	Accuracy ~90%, Sensitivity: 85%, Specificity: 92%	- Showcased the potential of deep learning for density categorization, with ResNet achieving high accuracy despite dataset limitations, but highlighted the need for larger datasets.

5	(Pullaiah et al., 2022)	- Analysed DDSM and MIAS datasets. - Enhanced datasets with augmentation methods. - Customized ResNet for cancer classification tasks. - Balanced performance and computational efficiency.	Customized ResNet	Accuracy, Sensitivity, Specificity	Accuracy ~90%, Sensitivity >85%	- Demonstrated that customized ResNet models can efficiently handle breast cancer classification with balanced accuracy and computational efficiency, emphasizing augmentation techniques.
6	(Daniel López-Cabrera et al., 2020)	- Utilized MIAS dataset with extensive augmentation. - Focused on two-stage classification approach. - Applied Inception v3 with transfer learning. - Emphasized distinguishing benign and malignant lesions.	ResNet50, InceptionV3	AUC, Accuracy, Sensitivity, Specificity	AUC: 93.3%, Accuracy: 91.3%, Sensitivity: 88.6%, Specificity: 98%	- Two-stage Inception v3-based classification achieved remarkable diagnostic accuracy, highlighting the importance of transfer learning for limited datasets.
7	(Khanna et al., 2022)	- Focused on hybrid approaches for BI-RADS classification. - Addressed dataset variability with robust preprocessing.	ResNet-18, DenseNet201, MLP	Accuracy, F1-score, Precision, Recall	High accuracy, balanced precision-recall trade-off	- Hybrid deep learning approaches effectively combined traditional ML and CNNs, providing flexible solutions for complex classification tasks.

		- Integrated traditional ML and deep learning techniques. - Highlighted flexibility of hybrid methods.				
8	(Shi et al., 2019)	- Addressed challenges of small datasets. - Pre-processed with scaling, normalization, and augmentation. - Utilized CNN architectures for density classification. - Proposed strategies for enhancing model generalization.	Lightweight CNNs	Accuracy, Sensitivity, Specificity	High accuracy (>85%) on small datasets	- Showed that preprocessing and augmentation can significantly enhance performance for small datasets, demonstrating the adaptability of custom CNNs.
9	(Wang, 2024)	- Employed diverse datasets for classification tasks. - Focused on preprocessing to enhance data quality. - Used CNN-based architectures tailored for images.	ResNet50, InceptionV3	AUC, Accuracy, Precision, Recall	AUC > 0.90, High accuracy	- Reinforced the utility of CNN-based approaches in cancer detection with high accuracy and AUC, highlighting preprocessing and hyperparameter tuning as critical steps.

		- Optimized hyperparameters to boost performance.				
10	(Gao et al., 2023)	- Compared traditional ML with DL for mammography. - Highlighted preprocessing and feature engineering methods. - Discussed applicability of CNNs for medical imaging. - Explored potential of ensemble learning and transfer learning.	ResNet50, InceptionV3	Accuracy, Precision, Recall	Accuracy: ~85%, High F1-score	- Demonstrated that both traditional ML and DL are effective for mammography, but CNNs provide superior diagnostic accuracy, especially when coupled with transfer or ensemble learning.

2.5 Deep Learning Algorithms

Deep learning represents a revolutionary technology in the diagnosis of breast cancer, especially when it has provided substantial improvements compared to traditional machine learning techniques (Macit & Sabanci, 2023; Pawar et al., 2022). Deep learning technology also facilitates automatic feature extraction from images, thereby overcoming the constraints associated with manual feature extraction in conventional machine learning (Liu et al., 2024). For instance, deep learning techniques such as Inception and ResNet50 are examined for breast cancer diagnosis utilising mammograms to underscore their efficacy in breast cancer categorisation (Ashwini et al., 2024). Across the study, convolutional neural networks (CNNs) has been used in most of the study since it's excel at the automatic extraction of spatial information from images, which is essential in medical imaging for the identification of patterns such as tumours, calcifications, or lesions (Pawar et al., 2022; Rahman et al., 2023). CNNs also widely use since it's capable of processing extensive and complex datasets with significant variation, including mammograms from diverse demographics, varying magnification levels, and differing imaging settings. Their versatility renders them an optimal answer for medical imaging challenges (Ashwini et al., 2024; Rahman et al., 2023).

Some of the examples of CNNs architecture that primary used in Residual Network (ResNet), Dense convolutional Network (DenseNet), and Inception. These architectures are primarily used in CNNs since it's capable of processing extensive and intricate datasets with significant heterogeneity, including mammograms from diverse demographics, varying magnification levels, and differing imaging settings. Their versatility renders them an optimal answer for medical imaging challenges (Ashwini et al., 2024; Liu et al., 2024). These architectures also exhibit robust performance by optimising parameter use and feature

extraction, particularly in skewed datasets (Sadad et al., 2021). Some of the architectures variants that has typically used is DenseNet201, InceptionV3, ResNet50, ResNet34. These variants have high accuracy which mostly above 85% due to its intricate interconnections, facilitating enhanced gradient propagation and feature reutilization, even in skewed datasets (Sadad et al., 2021) and ascribed to its uncomplicated, systematic design and transfer learning functionalities (Liu et al., 2024). The architectures employ supervised learning, since it successfully utilises labelled data to train models for specific classification tasks, rendering it the optimal alternative for breast cancer diagnosis (Liu et al., 2024). CNNs architectures thrive in supervised tasks as they are engineered to precisely correlate input data (e.g., mammograms) with specified outputs. Supervised learning guarantees excellent precision in categorising photos into established classifications by directly learning from annotated instances. This is crucial in medical imaging, as accurate predictions might influence patient outcomes (Ashwini et al., 2024; Rahman et al., 2023).

2.5.1 Residual Network (ResNet)

ResNet, a variant of deep convolutional neural networks, address the training challenges associated with extremely deep architectures. The fundamental principle of ResNet is to learn indirectly via residual connections, which focus on learning residual mappings rather than directly mapping from one layer to the subsequent layer. Consequently, there may be skip connections that circumvent the gradient flow during the training process, involving one or more layers that establish the "skip connection" (Pullaiah et al., 2022). It sustains a high level of accuracy regardless of the network's growing height. The essential residual connections provide gradient propagation during model training across numerous layers, potentially hundreds or thousands, which compromises performance due to optimisation complexity (Wang, 2024). The primary variables enhancing the convergence of deep networks have

rendered them more effective in executing complicated tasks such as image recognition, which frequently proved inadequate with conventional deep networks.

The different variations of the ResNet architecture are defined by their layer count, with ResNet-18, ResNet-34, ResNet-50, ResNet-101, and ResNet-152 being the most prevalent. The primary feature distinguishing these versions is the network's depth. Deeper networks typically exhibit superior performance on complicated tasks; however, they necessitate greater processing resources (Pullaiah et al., 2022). The training and optimisation of ResNet architectures involve transfer learning and pre-trained models as strategies to enhance performance while reducing training time (Khanna et al., 2022). Additionally, data augmentation, batch normalisation, and Adam optimisation are frequently employed to improve accuracy and mitigate overfitting (Wang, 2024). Consequently, ResNet has been extensively utilised in deep learning. Figure 4 show the Resnet 18 model that have customization for malignancy detection.

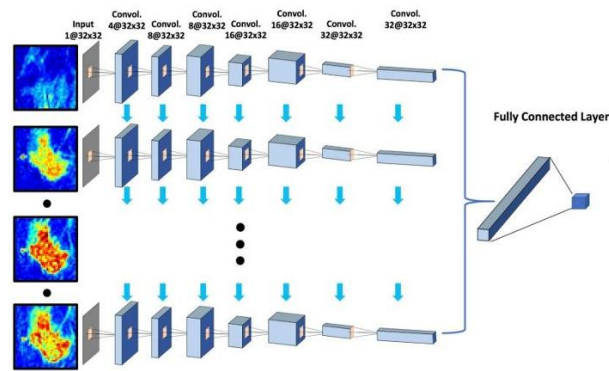


Figure. 1 Architecture of the proposed cRN-CNN method

Figure 2.2 Architecture of ResNet-18 that has been customized

2.5.2 Dense Convolutional Network (DenseNet)

DenseNet, is a deep learning architecture designed to enhance information transfer by establishing high-density connections between each layer and all preceding layers. In contrast to conventional convolutional networks, where each layer's inputs derive solely from the preceding layer, a DenseNet architecture allows every layer to receive inputs from all preceding layers, thereby facilitating the reuse of feature maps from all prior layers (Daniel López-Cabrera et al., 2020). This extensive connectivity has mitigated the vanishing gradient problem and enhanced efficiency in feature reuse, hence augmenting the capacity to learn with fewer parameters (Daniel López-Cabrera et al., 2020). The fundamental premise of DenseNet is concatenation, as opposed to addition utilised in residual connections. This architecture renders DenseNet highly efficient in training deep models, as the established dense connectivity facilitates robust gradient flow across all layers, resulting in improved performance and accelerated convergence (Wang, 2024). Optimal complexes typically involve tasks requiring highly efficient feature learning.

DenseNets differ in layer count and depth, including DenseNet-121, DenseNet-169, DenseNet-201, and DenseNet-264, which are distinguished by the quantity of dense blocks and the number of channels within each block. It has also been observed that deeper networks acquire highly complex properties but are less efficient in terms of computational resources required (Khanna et al., 2022). In numerous DenseNet applications, transfer learning is utilised, wherein pre-learnt weights from models previously trained on big datasets like ImageNet are refined for a particular job. Additionally, techniques like as batch normalisation, dropout, and weight decay are employed with various optimisation methods in this network.

2.5.3 Inception

The Inception model, a CNN architecture created by Google, has garnered considerable interest for its efficacy in deep learning applications, especially in picture classification, owing to its efficiency and superior performance (Paper 1). Inception belongs to the Inception model family, developed to address the difficulty of sustaining computing efficiency while attaining high accuracy (Paper 1). It utilizes a distinctive design in which many convolutional filters of different sizes are concurrently applied inside the same layer, referred to as "Inception modules" (Paper 2). This parallel processing allows the model to acquire a varied array of characteristics at many spatial resolutions, essential for intricate picture identification applications such as mammography analysis in breast cancer screening (Paper 3). The Inception network's capacity to adapt to diverse visual structures has rendered it a formidable contender in medical image classification tasks, including the differentiation between benign and malignant lesions in mammography images (Paper 4).

In deep learning applications, Inception is often optimized by transfer learning methods, particularly when the dataset is limited (Paper 2). The InceptionV3 model, pre-trained on extensive datasets such as ImageNet, may be customized for individual applications by altering the final layers of the network to correspond with the number of classes in the intended classification issue (Paper 3). This methodology has shown efficacy in breast cancer diagnosis, whereby the model, post fine-tuning, can categorize mammography pictures into classifications such as normal, benign, and malignant with considerable accuracy (Paper 4). Research indicates that Inception attains significant performance enhancements, achieving an accuracy of 96.67% on the DDSM dataset and 95.5% on the INbreast dataset, so demonstrating its resilience in medical image analysis (Paper 2). Moreover, the model's computational efficiency, realized via

the Inception module's parallel processing, reduces training duration while preserving competitive performance relative to other models such as ResNet and DenseNet (Paper 3).

The design of Inception diminishes computational expenses while enhancing the generalization capacity of deep learning models by allowing the network to assimilate varied information across various sizes (Paper 4). Its efficacy in breast cancer classification is well-documented, yielding remarkable outcomes in both binary and multi-class classifications (Paper 2). Inception enhances early and precise breast cancer diagnosis by feature extraction from mammograms, which is vital for improving patient outcomes (Paper 1). The model's ongoing refinement and optimization may improve the clinical efficacy of computer-aided detection (CAD) devices in standard mammography screening (Paper 4).

2.6 Evaluation Metrics

2.6.1 Accuracy

Accuracy, defined as the ratio of correct classifications to the total number of images, serves as an indicator of system reliability for the architecture comprising three classes (H. Liu et al., 2021). Accuracy quantifies the ratio of correct predictions generated by the model. The standard formula for determining accuracy is (Ramadhani et al., 2021):

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

Equation 2.1 Equation of Accuracy

Where,

TP = True Positive. The number of correctly predicted positive instances

TN = True Negative. The number of correctly predicted negative instances

FP = False Positive. The number of incorrectly predicted positive instances

FN = False Negative. The number of incorrectly predicted positive instances

The assessed performance parameters include fitness value, AUC, classification accuracy, sensitivity, and specificity (Sabani et al., 2022). Accuracy is a crucial parameter for assessing the classification efficacy of the suggested lightweight networks in managing tiny datasets successfully (Gandomkar et al., 2019). The accuracy metrics derived from the networks for various classification tasks are provided to evaluate the efficacy of BI-RADS classification (Pullaiah et al., 2022). We utilise cross-validation to evaluate the network's performance... Metrics such as Accuracy, Precision, and Recall are assessed to evaluate the model's efficacy (Daniel López-Cabrera et al., 2020). The training phase accuracy for the training set was approximately 95%, while for the validation set it was 92.6%. Nonetheless, the network attained an accuracy of 86.05% on the test set (H. Liu et al., 2021). The proposed methodology attained an accuracy of 87.5% with KNN employing feature selection, in contrast to 79.54% without feature selection (Sabani et al., 2022).

2.6.2 Area Under the Receiver Operating Characteristic Curve (AUC-ROC)

The AUC assesses the efficacy of features and classifiers across different threshold configurations (Khanna et al., 2022). The AUC is generally determined by graphing the true positive rate vs the false positive rate at various threshold levels and subsequently calculating the area beneath this curve (Ramadhani et al., 2021). The efficacy of predictive models was assessed by receiver operating characteristic (ROC) curve analysis and the area under the curve (AUC) (Sabani et al., 2022). Employing the stratified 10-fold cross-validation technique, the

KNN classifier attained an AUC of 0.82, underscoring the method's efficacy in differentiating between BI-RADS classes 3 and 4 (Gandomkar et al., 2019). The integrated deep learning model attained an AUC of 0.910 on the testing dataset, surpassing other models and illustrating its diagnostic efficacy (Pullaiah et al., 2022). The formula for AUC-ROC is:

$$AUC = \int_0^1 TPR(x)dx$$

Equation 2.2 Equation for AUC-ROC

where,

AUC = Area under the curve

TPR = True Positive Rate

2.6.3 Sensitivity (Recall)

Recall or known as sensitivity measures the ratio of true positive predictions to the total number of actual positive cases in the dataset (H. Liu et al., 2021). In clinical environments, sensitivity is essential for accurately identifying positive cases, hence reducing the likelihood of missed diagnoses (Ramadhani et al., 2021). Sensitivity, in conjunction with specificity and other metrics, is employed to assess the efficacy of binary classification systems, particularly in breast cancer screening (Sabani et al., 2022). The model's sensitivity indicates its capacity to accurately identify malignant instances, which is essential in medical diagnostics (Gandomkar et al., 2019). The suggested model attained a sensitivity of 85.3% on the testing dataset, indicating its robust capacity to detect malignant instances (Pullaiah et al., 2022). Employing the feature selection method, KNN attained a sensitivity of 88.88%, underscoring the

technique's accuracy in identifying affirmative situations (Daniel López-Cabrera et al., 2020).

The formula for sensitivity is:

$$Sensitivity = \frac{TP}{TP + FN}$$

Equation 2.3 Equation for Sensitivity

where,

TP = True Positive

FN = False Negative

2.6.4 Specificity

Specificity assesses a model's capacity to accurately detect negative situations, hence minimising false positives and superfluous medical procedures (H. Liu et al., 2021). Mammography must prevent patients from being unnecessarily subjected to additional invasive testing, hence reducing false positives in breast cancer screening (Ramadhani et al., 2021). The specificity of predictive models indicates their accuracy in differentiating between benign and malignant tumours, which is essential in diagnostic applications (Sabani et al., 2022). The suggested deep learning model attained a specificity of 91.9% in the testing dataset, indicating its proficiency in reliably classifying non-cancerous cases (Gandomkar et al., 2019). Through the feature selection method, KNN attained a specificity of 86.88%, demonstrating enhanced efficacy in eliminating false positives (Pullaiah et al., 2022). The formula for specificity is:

$$Specificity = \frac{TN}{TN + FP}$$

Equation 2.4 Equation for Specificity

where,

TN = True Negative

FP = False Positive

2.7 Tools

The deep learning tools and libraries in Python was primarily being used in the research paper for predicting the mammogram images. Here's a summary of the tools used in the research paper:

Table 2.2 Summary of the tools used in the research paper

Tool	Purpose	Reference
Keras	Training, validating, and testing the deep learning network using backpropagation with TensorFlow as backend.	(H. Liu et al., 2021)
Scikit-learn	Developing the clinical model for logistic regression analysis.	(Ramadhani et al., 2021).
TensorFlow	Backend for Keras to execute the deep learning model, ensuring computational efficiency.	(H. Liu et al., 2021)
Numpy	Performing numerical operations for preprocessing, enabling efficient matrix computations and data management.	(Sabani et al., 2022)
Pandas	Structuring and preprocessing the dataset, facilitating manipulation and cleansing of medical data.	(Gandomkar et al., 2019)
Matplotlib	Displaying accuracy and loss curves for model evaluation.	(Pullaiah et al., 2022)
Seaborn	Creating correlation matrices to examine feature interactions.	(Pullaiah et al., 2022)

OpenCV	Preparing mammography images, including resizing, segmentation, edge detection, and bounding box computations.	(Daniel López-Cabrera et al., 2020)
--------	--	-------------------------------------

2.8 Summary

This chapter explores the application of deep learning to enhance the diagnosis and classification of breast cancer through mammographic imaging. It emphasises the Breast Imaging Reporting and Data System (BI-RADS) for categorisation and addresses difficulties such as diversity in radiologist interpretations. This chapter examines multiple deep learning models, such as ResNet, DenseNet, and Inception, with a focus on preprocessing methods, dataset utilisation, and model parameterisation. Notable findings encompass the enhanced diagnostic accuracy, sensitivity, and specificity of deep learning models relative to conventional approaches. Challenges like data diversity, model generalisation, and computational resources are recognised, accompanied by suggestions for the integration of hybrid models and transfer learning methodologies to achieve optimal performance.

Chapter 3: Methodology

3.1 Introduction

This study adopts the Cross-Industry Standard Process for Data Mining (CRISP-DM) as shown in Figure 6, as a methodological framework that provides a structured and iterative approach to solving data-driven research problems. In this work, three CCN architectures, ResNet, DenseNet and Inception, are used to study their effectiveness in classifying mammography images based on the BI-RADS system. The dataset obtained from Kaggle serves as the basis for training and evaluating these models. By comparing their performance against rigorous evaluation metrics, this study aims to identify the most suitable deep learning architecture for accurate and robust BI-RADS classification, thereby contributing to improved diagnostic precision in clinical settings.

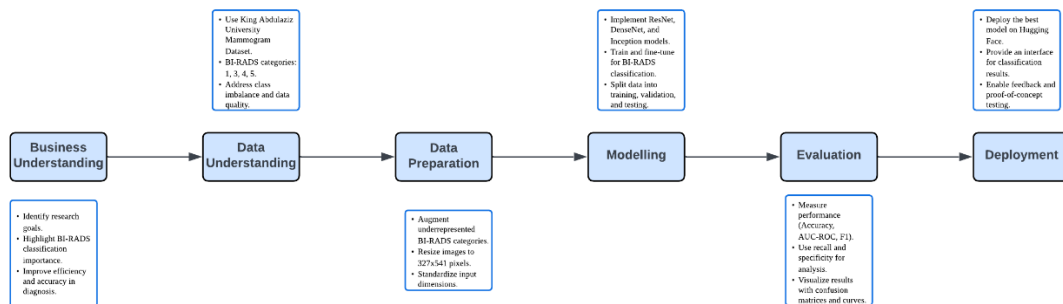


Figure 3.1 Model of CRISP-DM

3.2 Business Understanding

The BI-RADS classification system, which provides a standardized method for classifying mammography results, is critical for the diagnosis of breast cancer. These classifications help radiologists assess the likelihood of cancer and select the best approach to patient care. However, manual interpretation of mammography can be laborious and prone to

human error, which can lead to inconsistent diagnoses. The aim of this study is to increase the efficiency and accuracy of BI-RADS classification by using deep learning models, especially CCNs. This process could be automated to improve clinical workflow, reduce diagnostic variability, and help radiologists make faster and more accurate decisions about patient care.

Additionally, the ability to use state-of-the-art deep learning techniques to automate and improve BI-RADS classification addresses a few important healthcare needs. Because earlier intervention and therapy is possible, early and accurate detection of breast cancer is directly linked to better treatment outcomes for the patient. Solutions that improve diagnostic procedures are having a major impact on clinical practice and healthcare systems as the healthcare sector increasingly adopts AI and machine learning technologies. To help clinicians improve their diagnostic skills and overall patient care, this project aims to contribute to this changing landscape by shedding light on which deep learning model - ResNet, DenseNet or Inception - for the BI-RADS classification is most appropriate.

3.3 Domain Understanding

This study utilizes the King Abdulaziz University Mammogram Dataset obtained via Kaggle. The dataset comprises 2,378 mammography pictures, encompassing BI-RADS categories 1, 3, 4, and 5, essential for the identification and classification of breast tissue anomalies. The images differ in size, illustrating the heterogeneity present in mammography data from the real world. Each BI-RADS category denotes a distinct level of abnormality or danger, with BI-RADS 1 indicating normal findings, and BI-RADS 3, 4, and 5 reflecting escalating levels of suspicion for cancer. The dataset demonstrates class imbalance, characterized by an unbalanced distribution of images among the BI-RADS categories. This

mismatch presents a difficulty for model training, as categories with fewer data may result in skewed predictions or underrepresentation in categorization.

An in-depth examination of data quality indicates that the mammography pictures are highly appropriate for diagnostic purposes and deep learning applications. The photos demonstrate high resolution, capturing intricate anatomical details including tissue structure and microcalcifications. The contrast quality is suitable, allowing for unambiguous distinction between dense and fatty breast tissues, with margins and the skin line distinctly apparent. The dataset exhibits negligible noise and lacks substantial artifacts, including motion blur or extraneous objects, so guaranteeing that the photos are crisp and devoid of distortions. The field of view (FOV) is effectively handled, with the breast region occupying most of the image space, and appropriate anatomical coverage, including both craniocaudal (CC) and mediolateral oblique (MLO) views, is visible. Moreover, the homogeneity in labelling and exposure throughout the dataset guarantees consistency, hence enhancing model training and analysis.

Notwithstanding the generally high quality, certain possible issues are noted. Dense tissue patches in certain photos may mask anomalies, diminishing visibility in specific locations, although modest compression distortions are occasionally seen. Nonetheless, these flaws do not substantially impair the diagnostic quality of the dataset. The mammography dataset is comprehensive and dependable, offering high-resolution pictures with uniform quality across BI-RADS categories. The comprehensive and varied data, along with the issues of class imbalance, render this dataset an exemplary basis for the development and comparison of deep learning models for BI-RADS classification.

3.4 Data Preparation

This study reveals a significant class imbalance in the dataset, with BI-RADS 1 comprising 1,865 photos and BI-RADS 5 containing merely 24 images, resulting in a difference that could bias the deep learning models in favor of the majority class. Data augmentation techniques are employed to artificially enlarge the dataset, specifically targeting the underrepresented BI-RADS categories to rectify this imbalance. Augmentation procedures encompass rotation, horizontal flipping, zooming, and shifting, which provide novel versions of the original images while maintaining their inherent attributes. These strategies not only equilibrate the dataset but also enhance model generalization by presenting deep learning models with varied transformations of the mammography data.

To ensure uniformity in input across the three deep learning models, ResNet, DenseNet, and Inception all photos are downsized to a standardized dimension of 327x541 pixels. The scaling process is crucial as these models necessitate uniform input dimensions to execute convolution operations efficiently. Mammogram images in their native state exhibit diverse dimensions, and scaling guarantees homogeneity, facilitating seamless training and evaluation. Throughout this procedure, meticulous attention is devoted to preserving the aspect ratio and anatomical characteristics of the breast region to prevent distortion, which could undermine image integrity and impede model performance.

The integration of data augmentation and resizing guarantees that deep learning models are trained on a more comprehensive and representative dataset. Augmentation reduces the risk of overfitting by supplying models with varied examples, particularly for the underrepresented BI-RADS 3, 4, and 5 categories. Simultaneously, resizing guarantees computational economy and compatibility with model designs while maintaining the quality and diagnostic significance

of the mammograms. These preliminary measures are essential for improving the efficacy and dependability of deep learning models for BI-RADS classification.

3.5 Modelling

This study's modelling phase concentrates on the implementation and comparison of three advanced deep learning architectures ResNet, DenseNet, and Inception for the classification of mammograms into BI-RADS categories. These models were selected for their shown efficacy in picture classification challenges and their distinctive design principles, which enhance their performance in processing intricate visual data such as mammograms. Each model is executed utilising a pre-trained variation from renowned deep learning libraries (e.g., TensorFlow/Keras or PyTorch), employing transfer learning to tailor the models for the BI-RADS classification job. The dataset is partitioned into training, validation, and testing subsets to guarantee impartial performance assessment.

ResNet is known for its implementation of residual connections, which mitigate the vanishing gradient issue frequently seen in deep neural networks. ResNet facilitates robust training in deep architectures by permitting gradients to propagate directly across shortcut links. This work used a pre-trained ResNet-50 model, which provides an optimal balance of depth and computational efficiency. The model's capacity to discern intricate patterns in mammograms, including nuanced distinctions between dense and fatty tissues, renders it a formidable contender for BI-RADS classification. The implementation entails refining the end layers of ResNet to align its predictions with the BI-RADS categories, while the initial layers, trained on ImageNet, extract fundamental picture attributes.

DenseNet enhances feature reuse by interlinking each layer with all other layers in a feedforward manner. This dense connectivity improves feature propagation and decreases the parameter count, rendering DenseNet computationally efficient. This study utilises DenseNet-121 for its appropriateness in medical imaging applications. Its capacity to identify complex patterns in the dataset, including microcalcifications and minor tissue irregularities, is very beneficial for mammography analysis. During implementation, transfer learning is utilised by substituting the final classification layers with custom layers designed for the BI-RADS categories, followed by fine-tuning to enhance classification accuracy.

Inception is a more straightforward architecture than ResNet and DenseNet, distinguished by its implementation of consistent convolutional layers utilising modest 3x3 filters. Notwithstanding its simplicity, Inception has exhibited robust performance in picture classification tasks through the utilisation of depth and hierarchical feature extraction. Inception is chosen for this project due to its equilibrium between computational efficiency and depth. The design's simplicity facilitates the interpretation of learnt features, enhancing comprehension of the model's classification of BI-RADS categories. Implementation entails substituting the dense layers at the apex of the network with bespoke layers tailored for the classification purpose, succeeded by fine-tuning to acclimatise the pre-trained model to the mammography dataset.

3.6 Evaluation

The evaluation phase of this study is measuring the performance of the deep learning models such as ResNet, DenseNet, and Inception utilizing a set of rigorous evaluation measures specifically designed for the BI-RADS classification task. Accuracy functions as the principal

metric, denoting the ratio of correctly classified mammograms within all BI-RADS categories. Although accuracy serves as a broad indicator of model performance, it may be inadequate in isolation, especially in instances of class imbalance. Consequently, further metrics like AUC-ROC are utilised to assess the models' proficiency in differentiating between classes. The AUC-ROC score is vital for this research since it quantifies the balance between sensitivity and specificity at different classification thresholds, providing enhanced understanding of the models' diagnostic performance.

Recall and specificity are incorporated to further evaluate the models' performance, especially regarding their significance in medical diagnosis. Recall, or sensitivity, denotes the ratio of accurately detected positive cases (e.g., BI-RADS categories 3, 4, and 5), which is essential for reducing false negatives that may lead to overlooked diagnosis. Conversely, specificity quantifies the ratio of accurately detected genuine negatives, indicating the model's capacity to minimise false positives. These indicators are essential in a healthcare context, as they directly influence the accuracy of predictions and patient outcomes. The evaluation process integrates both recall and specificity, providing a balanced assessment of the models' strengths and weaknesses in detecting abnormalities.

The efficacy of ResNet, DenseNet, and Inception is methodically evaluated using the specified metrics, with findings displayed in tables and visual representations, including confusion matrices and AUC-ROC curves. This comparative analysis elucidates the advantages of each model, determining the optimal design for BI-RADS categorisation. Furthermore, training and validation accuracy and loss curves are illustrated to assess convergence and identify overfitting or underfitting throughout model training. The work employs a combination of evaluation metrics to conduct a thorough and rigorous assessment of the models, thereby

identifying the most dependable deep learning model for BI-RADS classification of mammograms.

3.7 Deployment

This study's deployment phase centres on showcasing the optimal deep learning model in a user-friendly style, prioritising demonstration above comprehensive clinical integration. The selected method entails deploying the trained model on a cloud-based platform for public engagement and assessment. This strategy enables academics, physicians, and other stakeholders to examine the model's capabilities and outputs in a controlled environment rather than implementing it directly in a clinical situation. The deployment is designed to offer a functioning interface for uploading mammography pictures and displaying the BI-RADS classification results, including confidence scores. This allows stakeholders to evaluate the model's performance and reliability without necessitating intricate local deployments or integrations.

This demonstration employs Hugging Face's model hub as the hosting platform. Hugging Face offers an efficient and accessible platform for demonstrating machine learning models, especially deep learning models, featuring integrated API support and an intuitive interface. The trained model, exported in a format compatible with Hugging Face (e.g., PyTorch or TensorFlow), will be uploaded and incorporated into the platform. A bespoke Gradio or Streamlit interface will enable users to upload mammography pictures and observe the model's predictions instantaneously. This configuration not only enables wider access and feedback from the research community but also acts as a proof-of-concept for possible future integration into clinical workflows.

3.8 Summary

The research utilises the CRISP-DM framework to systematically tackle the issues associated with BI-RADS classification in mammography, utilising a dataset sourced from Kaggle. Three deep learning models, ResNet, DenseNet, and Inception are employed, chosen for their shown efficacy in medical imaging. The dataset undergoes preprocessing via data augmentation and scaling methods to rectify class imbalance and assure consistency while preserving diagnostic accuracy. These measures improve the models' capacity to generalise and accurately identify mammography images within BI-RADS categories.

Model evaluation employs a thorough array of criteria, such as accuracy, AUC-ROC, recall, and specificity, to deliver a fair and rigorous assessment of diagnostic performance. Deployment emphasises the presentation of the optimal model via an intuitive interface on Hugging Face, enabling users to input mammography photos and obtain classification outcomes. This method acts as a proof-of-concept, illustrating the model's capability for incorporation into clinical processes while encouraging input from the scientific community.

Chapter 4: Implementation

4.1 Introduction

This chapter explores the comprehensive execution of the suggested mammographic picture categorisation system utilising deep learning methodologies. The implementation includes several essential phases: data preparation, model building, evaluation, and deployment. A systematic methodology was employed to address difficulties commonly encountered in medical imaging datasets, including class imbalance and fluctuation in image dimensions, so guaranteeing the training data was well-organised and favourable to efficient learning. The King Abdulaziz University Mammogram Dataset, particularly the mediolateral oblique (MLO) view images, was chosen for its clinical significance and extensive representation of breast tissue regions susceptible to cancers.

The research further investigates the adaptability and optimisation of three advanced convolutional neural network architectures—ResNet50, DenseNet121, and InceptionV3—via transfer learning. Each model was carefully tailored to categorise mammographic pictures based on BI-RADS classifications, utilising pre-trained weights and integrating domain-specific modifications. Evaluation criteria including accuracy, recall, specificity, F1-score, and AUC-ROC were utilised to thoroughly analyse the diagnostic performance of the models. The chapter concludes with the implementation of the most effective model on the Hugging Face platform, offering a practical interface for clinical use and future research development.

4.2 Data Preparation

The data preparation stage is essential to this study, establishing the groundwork for efficient training and assessment of deep learning models. The dataset employed is the King Abdulaziz University Mammogram Dataset, obtained from Kaggle, comprising

mammographic pictures annotated with BI-RADS classifications. To maintain consistency and eradicate redundancy, only mediolateral oblique (MLO) view photos were chosen for this project. MLO views are favoured in clinical environments because they provide enhanced visualisation of breast tissue, encompassing the pectoral muscle and upper outer quadrant, where cancers are more commonly found.

4.2.1 Class Imbalance and Offline Augmentation

A primary hurdle faced was class imbalance, a prevalent issue in medical datasets. BI-RADS 1 had a substantial number of samples, while BI-RADS 5, denoting a high suspicion of malignancy, was significantly under-represented. A stringent offline data augmentation method was employed to tackle this issue. In offline augmentation, additional picture variations are produced and preserved before training, rather than being generated dynamically during the training process. The objective was to standardise each class to 930 pictures, reflecting the maximum shared quantity among the categories post-augmentation.

4.2.2 Image Resizing Strategy

In addition, to guarantee uniform input dimensions for all models, all pictures were reduced to a standard resolution of 327×541 pixels. This dimension was chosen since it signifies the smallest form in all photos, notably corresponding with BI-RADS 5 samples, hence reducing quality degradation from upsampling and preserving anatomical structural integrity. Resizing was meticulously executed with the preservation of aspect ratio, including padding methods to prevent distortion that could misguide the model during feature extraction. The preprocessing pipeline was customised for each class according to its initial picture count. BI-RADS 1 images necessitated simply resizing; BI-RADS 3 and 4 images needed both resizing

and augmentation; but BI-RADS 5, due to its exceedingly low quantity (12 images), required significant augmentation exclusively.

The final class distribution after augmentation and resizing is as follows based on Table 1:

Table 4.1 Class Distribution before and after augmentation and resizing

BI-RADS Category	Original Image Count	Preprocessing	Final Image Count
BI-RADS 1	930	Resizing only	930
BI-RADS 3	196	Resizing and Augmentation	930
BI-RADS 4	54	Resizing and Augmentation	930
BI-RADS 5	12	Augmentation only	930

4.2.3 Resizing and Augmentation Functions

a) *resize_with_padding()* for BI-RADS 3 & 4

```
# Function to resize and pad images while maintaining aspect ratio
def resize_with_padding(image, target_width, target_height):
    # Get the current dimensions of the image
    height, width = image.shape[:2]

    # Calculate the aspect ratio of the image
    aspect_ratio = width / height

    # Compute new dimensions while maintaining the aspect ratio
    if aspect_ratio > target_width / target_height: # Wider image
        new_width = target_width
        new_height = int(target_width / aspect_ratio)
    else: # Taller image
        new_height = target_height
        new_width = int(target_height * aspect_ratio)

    # Resize the image
    resized_image = cv2.resize(image, (new_width, new_height))

    # Add padding to make it the target size (327x541)
    top = (target_height - new_height) // 2 if new_height < target_height else 0
    bottom = target_height - new_height - top if new_height < target_height else 0
    left = (target_width - new_width) // 2 if new_width < target_width else 0
    right = target_width - new_width - left if new_width < target_width else 0

    # Only apply padding if necessary
    if top >= 0 and bottom >= 0 and left >= 0 and right >= 0:
        padded_image = cv2.copyMakeBorder(resized_image, top, bottom, left, right, cv2.BORDER_CONSTANT, value=(0, 0, 0))
    else:
        padded_image = resized_image

    return padded_image
```

Figure 4.1 Function to resize and pad images

The function *resize_with_padding* based on Figure 4.1 ensures that each image fits within the target resolution of 327×541 pixels without distorting the original aspect ratio. It calculates appropriate scaling and adds black padding (zero-pixel values) where necessary. This

method was crucial for classes BI-RADS 3 and 4, where images had varying resolutions. Padding helped standardize dimensions across the dataset while maintaining anatomical features.

b) *center_crop_zoom*() for BI-RADS 1

```
# Function to crop the region of interest (ROI) and zoom into the breast area
def center_crop_zoom(image, target_width, target_height, crop_factor=0.7):
    # Get the current dimensions of the image
    height, width = image.shape[:2]

    # Calculate the size of the cropping box based on the crop factor
    crop_width = int(width * crop_factor)
    crop_height = int(height * crop_factor)

    # Compute the starting coordinates of the crop box to center the crop
    startx = (width - crop_width) // 2
    starty = (height - crop_height) // 2

    # Crop the image to zoom in
    cropped_image = image[starty:starty + crop_height, startx:startx + crop_width]

    # Resize the cropped image to the target size
    resized_image = cv2.resize(cropped_image, (target_width, target_height))

    return resized_image
```

Figure 4.2 Function to crop the ROI and Zoom area

The function *center_crop_zoom* was applied specifically to BI-RADS 1 images to slightly zoom into the region of interest while still maintaining contextual boundaries. This enhanced feature visibility in otherwise normal tissue while keeping the final size at 327×541. As shown in figure 4.2, the crop factor (e.g., 0.7) removes outer, potentially less informative pixels, improving the clarity of the breast region fed into the network.

c) *apply_augmentation* for BI-RADS 3,4 and 5

```

# Function to apply manual augmentation (rotation, shift, zoom, etc.)
def apply_augmentation(image):
    # Random rotation (between -10 and 10 degrees)
    rotation_angle = randint(-10, 10)
    M = cv2.getRotationMatrix2D((image.shape[1]//2, image.shape[0]//2), rotation_angle, 1)
    rotated_image = cv2.warpAffine(image, M, (image.shape[1], image.shape[0]))

    # Random width and height shift (between -5% and 5%)
    width_shift = randint(-5, 5) / 100.0 * image.shape[1]
    height_shift = randint(-5, 5) / 100.0 * image.shape[0]
    M = np.float32([[1, 0, width_shift], [0, 1, height_shift]])
    shifted_image = cv2.warpAffine(rotated_image, M, (rotated_image.shape[1], rotated_image.shape[0]))

    # Random zoom (between 0.95 and 1.05)
    zoom_factor = 1 + randint(-5, 5) / 100.0
    zoomed_image = cv2.resize(shifted_image, None, fx=zoom_factor, fy=zoom_factor)

    # Resize back to 327x541 to maintain consistency (this is the key change)
    target_width = 327
    target_height = 541
    resized_image = cv2.resize(zoomed_image, (target_width, target_height))

    return resized_image

```

Figure 4.3 Function to apply offline augmentation

For BI-RADS 3,4 and 5, the original image count was critically low. Therefore, extensive augmentation was required. The *apply_augmentation* function applies several transformations: random rotation, shifting, zooming, and finally resizes the output back to 327×541 to ensure uniform input shape. This augmentation approach (figure 4.3) artificially increases data variability, which is vital for avoiding overfitting on the limited original BI-RADS 5 data. The transformation parameters were randomized to simulate natural variations in mammographic imaging conditions.

Through this structured and class-sensitive preprocessing strategy, the final dataset achieves both balance and uniformity, creating an optimal foundation for training deep learning models capable of accurately classifying mammographic images across all BI-RADS categories.

4.3 Modelling

This study implemented and compared three state-of-the-art convolutional neural network architectures ResNet50, DenseNet121, and InceptionV3 for the classification of

mammographic images into BI-RADS categories. All models were developed using the transfer learning approach with pre-trained ImageNet weights. The implementation followed a two-phase training protocol:

1. Freezing the base layers and training only the top layers
2. Unfreezing a portion of the base network for fine-tuning using a lower learning rate.

This hybrid strategy leverages pre-learned visual features while allowing domain-specific adjustments to mammogram data. Each model was trained using a batch size of 16, categorical cross-entropy loss, and the Adam optimizer.

4.3.1 Modelling with DenseNet121

```
base_model = DenseNet121(weights='imagenet', include_top=False, input_shape=(327, 541, 3))
# base_model = ResNet50(weights='imagenet', include_top=False, input_shape=(327, 541, 3))
base_model.trainable = False

model = models.Sequential([
    base_model,
    layers.GlobalAveragePooling2D(),
    layers.Dense(512, activation='relu',
                 kernel_regularizer=regularizers.l2(weight_decay)),
    layers.Dropout(0.7),
    layers.Dense(4, activation='softmax',
                 kernel_regularizer=regularizers.l2(weight_decay))
])
```

Figure 4.4 DenseNet121 based model architecture

```

# --- Phase 1: Freeze base ---
history1 = model.fit(
    train_generator,
    steps_per_epoch=train_generator.samples // train_generator.batch_size,
    validation_data=val_generator,
    validation_steps=val_generator.samples // val_generator.batch_size,
    epochs=10,
    callbacks=callback_list
)

# --- Phase 2: Unfreeze last N layers ---
N = 20
for layer in base_model.layers[-N:]:
    layer.trainable = True

model.compile(optimizer=Adam(5e-6), loss='categorical_crossentropy', metrics=['accuracy'])

history2 = model.fit(
    train_generator,
    steps_per_epoch=train_generator.samples // train_generator.batch_size,
    validation_data=val_generator,
    validation_steps=val_generator.samples // val_generator.batch_size,
    epochs=55 - len(history1.history['loss']),
    callbacks=callback_list
)

```

Figure 4.5 Two phase training with fine tuning for DenseNet121

DenseNet121 is an architecture characterised by thick connections, wherein each layer obtains inputs from all prior layers. This approach improves feature propagation, promotes reuse, and results in superior performance with fewer parameters than conventional CNNs. This method involved initialising DenseNet121 using pre-trained ImageNet weights and appending a custom classification head (figure 4.4). The head comprised a global average pooling layer, a 512-unit dense layer with L2 regularisation (*weight_decay* = 1e-3), a dropout layer with a rate of 0.7, and a softmax output layer.

During the initial phase, the convolutional base was immobilised to preserve the acquired characteristics as illustrated in figure 4.5. The model was trained exclusively on the recently incorporated dense layers. During the second phase, the final 20 layers of the base model were unfrozen and fine-tuned with a diminished learning rate of 5e-6. This enabled the model to enhance intricate visual representations within the mammography dataset.

The architectural benefit of DenseNet its capacity to identify small irregularities like microcalcifications was particularly pertinent to mammography analysis. The evaluation encompassed macro-averaged recall, F1-score, class-level specificity, ROC-AUC, and confusion matrix analysis for a thorough assessment.

4.3.2 Modelling with InceptionV3

```
base_model = InceptionV3(weights='imagenet', include_top=False, input_shape=(327, 541, 3))
base_model.trainable = False

model = models.Sequential([
    base_model,
    layers.GlobalAveragePooling2D(),
    layers.Dense(512, activation='relu', kernel_regularizer=regularizers.l2(weight_decay)),
    layers.Dropout(0.6),
    layers.Dense(4, activation='softmax', kernel_regularizer=regularizers.l2(weight_decay))
])
```

Figure 4.6 InceptionV3 based model architecture

```
# --- Phase 1: Train with frozen base ---
history1 = model.fit(
    train_generator,
    steps_per_epoch=train_generator.samples // train_generator.batch_size,
    validation_data=val_generator,
    validation_steps=val_generator.samples // val_generator.batch_size,
    epochs=10,
    callbacks=callback_list
)

# --- Phase 2: Unfreeze top 20 layers ---
for layer in base_model.layers[-20:]:
    layer.trainable = True

model.compile(optimizer=Adam(5e-6), loss='categorical_crossentropy', metrics=['accuracy'])

history2 = model.fit(
    train_generator,
    steps_per_epoch=train_generator.samples // train_generator.batch_size,
    validation_data=val_generator,
    validation_steps=val_generator.samples // val_generator.batch_size,
    epochs=55 - len(history1.history['loss']),
    callbacks=callback_list
)
```

Figure 4.7 Two phase training with fine tuning for InceptionV3

InceptionV3 utilises a distinctive architecture using Inception modules that implement many convolutional filters of differing dimensions concurrently. This enables the model to

proficiently capture multi-scale picture information, which is especially advantageous in mammographic imaging where abnormalities may differ in size and context. The model was initialised using pre-trained ImageNet weights and equipped with a bespoke top layer based on figure 4.6. This layer configuration comprises a global average pooling layer, a 512-unit dense layer utilising ReLU activation and L2 regularisation, succeeded by a dropout layer (rate = 0.6), and a softmax layer for four-class classification.

During the initial step, the basic model was immobilised to utilise broad visual attributes from ImageNet. Only the bespoke classification head was trained. During fine-tuning, the upper 20 layers of the base were unfrozen and retrained at a reduced learning rate (5e-6) to specialise on domain-specific features (refer figure 4.7).

InceptionV3's capacity to represent both global context and intricate details rendered it a formidable contender for the BI-RADS classification assignment. The model was assessed using identical stringent criteria and visualisations as for ResNet50 and DenseNet121, ensuring consistency in performance evaluation.

4.3.3 Modelling with ResNet50

```
# --- Build model using ResNet50 ---
base_model = ResNet50(weights='imagenet', include_top=False, input_shape=(327, 541, 3))
base_model.trainable = False

model = models.Sequential([
    base_model,
    layers.GlobalAveragePooling2D(),
    layers.Dense(512, activation='relu'),
    layers.Dropout(0.5),
    layers.Dense(4, activation='softmax') # 4 classes: B1, B3, B4, B5
])
```

Figure 4.8 ResNet50 based model architecture

```

# --- Phase 1 ---
history1 = model.fit(
    train_generator,
    steps_per_epoch=train_generator.samples // train_generator.batch_size,
    validation_data=val_generator,
    validation_steps=val_generator.samples // val_generator.batch_size,
    epochs=10,
    callbacks=[checkpoint_cb, earlystop_cb]
)

# --- Phase 2 ---
for layer in base_model.layers[-30:]:
    layer.trainable = True

model.compile(optimizer=Adam(1e-5), loss='categorical_crossentropy', metrics=['accuracy'])

history2 = model.fit(
    train_generator,
    steps_per_epoch=train_generator.samples // train_generator.batch_size,
    validation_data=val_generator,
    validation_steps=val_generator.samples // val_generator.batch_size,
    epochs=55 - len(history1.history['loss']),
    callbacks=[checkpoint_cb, earlystop_cb]
)

```

Figure 4.9 Two-phase training with fine-tuning for ResNet50

ResNet50 (Residual Network) is a deep convolutional neural network that introduces residual learning to overcome vanishing gradients by enabling direct gradient propagation through shortcut connections. This makes it particularly effective for deep image classification tasks like BI-RADS categorization.

The architecture was initialized with pre-trained ImageNet weights, with the original classification head removed and replaced by a custom one (refer to figure 4.8). The new head consists of a global average pooling layer, a dense layer with 512 ReLU-activated neurons, a dropout layer with a rate of 0.5, and a softmax output layer with four units corresponding to the BI-RADS categories.

In the initial training phase, all convolutional layers from ResNet50 were frozen to preserve general image features learned from ImageNet (refer to figure 4.9). Only the custom top layers were trained. In the fine-tuning phase, the top 30 layers of the base model were

unfrozen to allow deeper feature adaptation specific to mammographic textures. A lower learning rate (1e-5) was used to avoid disrupting the pre-trained weights.

This two-step strategy ensured efficient convergence and reduced overfitting. Training progress and performance were monitored using early stopping and validation accuracy metrics, and final performance was evaluated through metrics including accuracy, recall, f1-score, specificity, confusion matrix, and ROC-AUC curves.

4.4 Evaluation

All models were evaluated on a reserved validation set using multiple performance metrics to ensure a comprehensive assessment of classification capabilities. Metrics included accuracy, AUC-ROC, recall, f1-score, specificity, confusion matrix, and training dynamics such as loss and accuracy over epochs. These metrics were computed using *scikit – learn* and visualized using *matplotlib* and *seaborn*. The following subsections describe each metric and how they were implemented and interpreted.

4.4.1 Accuracy

```
# --- Compute Specificity ---
def compute_specificity(cm):
    specificity_list = []
    for i in range(len(cm)):
        TP = cm[i, i]
        FP = cm[:, i].sum() - TP
        FN = cm[i, :].sum() - TP
        TN = cm.sum() - (TP + FP + FN)
        specificity = TN / (TN + FP) if (TN + FP) > 0 else 0
        specificity_list.append(specificity)
    return specificity_list

recall = recall_score(y_true, y_pred, average=None)
f1 = f1_score(y_true, y_pred, average=None)
specificity = compute_specificity(cm)
support = cm.sum(axis=1)
accuracy = np.mean(y_true == y_pred)

report = classification_report(y_true, y_pred, target_names=class_names, output_dict=True)
avg_recall = recall_score(y_true, y_pred, average='macro')
avg_f1 = f1_score(y_true, y_pred, average='macro')
avg_spec = np.mean(specificity)
total_support = np.sum(support)

# --- Formatted Classification Report ---
print("\nClassification Report:\n")
print(f'{"":<12}{"recall":<12}{"f1-score":<12}{"specificity":<16}{"support":<10}')
for i, label in enumerate(class_names):
    print(f"[label]:<12}{recall[i]:<12.4f}{f1[i]:<12.4f}{specificity[i]:<16.4f}{support[i]:<10}")

print(f"\n{'accuracy':<12}{'':<24}{accuracy:<16.4f}{total_support:<10}")
print(f'{"macro avg":<12}{avg_recall:<12.4f}{avg_f1:<12.4f}{avg_spec:<16.4f}{total_support:<10}')
print(f'{"weighted avg":<12}{report["weighted avg"]['recall']:<12.4f}{report["weighted avg"]['f1-score']:<12.4f}{avg_spec:<16.4f}{total_support:<10}')
```

Figure 4.10 Performance Metric for all three models

Accuracy measures the proportion of correct predictions made by the model on the validation dataset. It quantifies the overall effectiveness of the classifier by calculating how many images were correctly categorized into their respective BI-RADS classes relative to the total number of images evaluated. While accuracy provides a straightforward and intuitive indicator of performance, it is limited in contexts where the dataset classes are imbalanced—as is common in medical imaging—because a high accuracy can be artificially inflated by the dominant class.

In this study, accuracy was computed directly from the predicted class labels and true labels obtained from the model’s output on the validation set. This calculation was part of a comprehensive evaluation block, which also included metrics more sensitive to class imbalance and clinical relevance (refer to figure 4.10). This integrated approach ensured that accuracy was interpreted alongside other metrics that reveal subtler aspects of model performance, such as recall and specificity.

4.4.2 F1-Score

The F1-score provides a balance between precision (the accuracy of positive predictions) and recall (the coverage of actual positives) and is particularly valuable when class distribution is uneven. In medical diagnostics, an elevated F1-score reflects that the model not only detects a high proportion of true positives but also minimizes false alarms, a desirable characteristic in breast cancer screening where both missed detections and false positives have significant clinical consequences.

In this study, F1-scores were computed per BI-RADS class to capture performance variability among different severity levels, and a macro-averaged F1-score was calculated to give equal weight to each class, mitigating the effect of imbalance. This comprehensive

calculation is included in the shared evaluation code block based on figure 4.10, ensuring all related metrics stem from the same prediction data, thereby facilitating integrated analysis.

4.4.3 Recall

Recall, or sensitivity, is a critical metric in breast cancer diagnosis as it quantifies the proportion of actual positive cases correctly identified by the model. High recall reduces the chance of false negatives, which could lead to missed cancers and delayed treatment. Therefore, achieving strong recall is paramount for ensuring patient safety and effective clinical outcomes.

The study calculated recall for each BI-RADS category individually, as well as a macro-averaged recall score to assess the model's ability to detect abnormalities consistently across all classes. By computing recall within the consolidated evaluation block (refer to figure 4.10), the study ensured uniformity of inputs and outputs, enabling precise comparisons with other performance indicators such as specificity and F1-score.

4.4.4 Specificity

Specificity measures the model's ability to correctly identify negative cases, that is, benign or normal mammograms, minimizing false positive results. High specificity reduces unnecessary biopsies and patient anxiety stemming from overdiagnosis. This metric complements recall, providing a balanced perspective on the model's diagnostic accuracy.

In this work, specificity was computed from the confusion matrix using a dedicated function that calculates true negatives and false positives for each class. This approach allowed a per-class specificity assessment, as well as an average value summarizing overall model performance in rejecting non-cancer cases. The specificity calculation is part of the shared evaluation code segment presented in figure 4.10, ensuring it was computed alongside other critical metrics from consistent prediction outputs.

4.4.5 ROC-AUC

```
# --- ROC AUC Curve ---  
y_true_bin = label_binarize(y_true, classes=list(range(len(class_names))))  
roc_auc = roc_auc_score(y_true_bin, val_preds, average='macro', multi_class='ovr')
```

Figure 4.11 ROC-AUC score calculation for multi-class classification

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) assesses the model's ability to distinguish between different BI-RADS classes over a range of classification thresholds. Unlike accuracy, which considers only one fixed threshold, AUC-ROC reflects performance across all possible trade-offs between sensitivity and specificity. In the context of mammogram classification, a high AUC indicates that the model is capable of effectively separating normal, benign, and malignant tissue characteristics, which is critical for reducing diagnostic errors.

This metric was calculated using a one-vs-rest strategy, where each BI-RADS class is treated as the positive class in turn, and the remaining classes are grouped as negative. The resulting ROC curves were then aggregated using a macro-average to summarize performance across all classes fairly. The code to calculate and visualize AUC-ROC is part of the unified evaluation block shared by multiple metrics, as shown in Figure 4.11. This integrated implementation promotes consistency in evaluation and aids comparative interpretation.

4.4.6 Confusion Metrix

```
# --- Confusion Matrix ---  
cm = confusion_matrix(y_true, y_pred)  
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=class_names, yticklabels=class_names)  
plt.title('Confusion Matrix')  
plt.xlabel('Predicted')  
plt.ylabel('True')  
plt.show()
```

Figure 4.12 Confusion matrix visualization using heatmap

The confusion matrix offers an intuitive and detailed breakdown of the classifier's predictions across all BI-RADS categories. Each row of the matrix corresponds to the true class labels, while each column represents the predicted classes. This matrix enables identification of specific patterns of misclassification, such as confusion between adjacent BI-RADS levels that may share similar radiological features.

In this study, the confusion matrix was visualized as a heatmap to clearly present the counts of true positives, false positives, true negatives, and false negatives for each class (refer to figure 4.12). This visualization is critical for diagnosing model weaknesses and for guiding future improvements, especially in categories where clinical differentiation is inherently challenging. The plotting code ensures that class names label the axes for ease of interpretation.

4.4.7 Accuracy and Loss over Epochs

```
# --- Accuracy/Loss Plots ---
plt.figure(figsize=(12, 4))
plt.subplot(1, 2, 1)
plt.plot(history['accuracy'], label='Train Acc')
plt.plot(history['val_accuracy'], label='Val Acc')
plt.title('Accuracy over Epochs')
plt.xlabel('Epochs')
plt.ylabel('Accuracy')
plt.legend()

plt.subplot(1, 2, 2)
plt.plot(history['loss'], label='Train Loss')
plt.plot(history['val_loss'], label='Val Loss')
plt.title('Loss over Epochs')
plt.xlabel('Epochs')
plt.ylabel('Loss')
plt.legend()
```

Figure 4.13 Training and validation accuracy and loss plots

Tracking training and validation accuracy and loss across epochs is crucial for evaluating the model's learning and generalisation capabilities. These graphs illustrate the

advancement of the learning process, facilitating the identification of problems such as underfitting or overfitting. The training accuracy and loss curves demonstrate the model's performance on the training dataset as it processes additional samples, whereas the validation curves reflect the model's ability to generalise to novel data. An optimal learning method demonstrates enhanced accuracy and diminished loss for both training and validation datasets, exhibiting little divergence.

In this study, the curves were generated utilising the aggregated training histories from both training stages (frozen base and fine-tuning), as shown in Figure 4.13. The analysis of these plots facilitated the identification of ideal stopping locations and the efficacy of regularisation techniques, including dropout and early stopping callbacks.

4.5 Deployment

The deployment phase demonstrates the practical application of the trained deep learning model in a user-friendly interface, enabling end users such as clinicians or researchers to utilize the BI-RADS classification system without requiring technical knowledge of model internals. This section outlines the deployment platform and explains the design and operation of the application script (*app.py*) hosted on the Hugging Face platform.

4.5.1 Deployment Platform

The trained DenseNet121 model was deployed on the Hugging Face Model Hub, a popular cloud platform facilitating easy sharing and hosting of machine learning models. Hugging Face offers integrated API services and streamlines interfaces like Gradio, allowing rapid development of interactive web apps without complex infrastructure.

Using Hugging Face for deployment brings several benefits:

1. Accessibility: Users can upload mammogram images directly through a web interface.

2. Scalability: The platform handles backend compute resources, scaling as necessary.
3. Integration: API support allows for further extension or embedding in clinical systems.
4. Rapid prototyping: Minimal setup is needed, making the model publicly accessible for testing and feedback.

The deployment emphasizes demonstration and usability over full clinical integration, serving as a proof-of-concept that bridges model development and real-world application.

4.5.2 Application Interface Key Components

The deployment application script *app.py* leverages Gradio to create an intuitive interface for the BI-RADS classification model. Below are three critical code sections that underpin its functionality, each accompanied by detailed explanations and references to corresponding figures.

a. Model Loading and Input Size Definition

```
# Load your saved model (change path if needed)
model_path = "final_densenet_savedmodel"
model = tf.keras.models.load_model(model_path)

# Your model input size
IMG_HEIGHT = 327
IMG_WIDTH = 541
```

Figure 4.14 Loading saved DenseNet model

This snippet initializes the deployment by loading the pre-trained DenseNet121 model saved in TensorFlow's format (refer to figure 4.14). Defining the input size explicitly matches the dimensions used during model training, ensuring consistency and proper operation during inference.

b. Image Preprocessing Function

```
def predict(image: Image.Image):
    img_array = preprocess(image)
    preds = model.predict(img_array)[0] # get first batch
    # Return dictionary of class probabilities ONLY
    results = {class_names[i]: float(preds[i]) * 100 for i in range(len(class_names))}
    return results
```

Figure 4.15 Prediction function for class probabilities

The preprocess function prepares user-uploaded mammogram images for model prediction (refer to figure 4.15). It converts images to RGB format, resizes them to the fixed input dimensions, normalizes pixel values to the [0, 1] range, and adds a batch dimension. This processing pipeline ensures compatibility with the DenseNet input requirements.

c. Prediction and Gradio Interface Setup

```
def predict(image: Image.Image):
    img_array = preprocess(image)
    preds = model.predict(img_array)[0] # get first batch
    # Return dictionary of class probabilities ONLY
    results = {class_names[i]: float(preds[i]) * 100 for i in range(len(class_names))}
    return results

# Gradio UI
title = "BI-RADS Classification (DenseNet121)"
description = "Upload a mammogram image and get predicted BI-RADS classification with probabilities."

iface = gr.Interface(
    fn=predict,
    inputs=gr.Image(type="pil"),
    outputs=gr.Label(num_top_classes=4),
    title=title,
    description=description,
    allow_flagging="never"
)
```

Figure 4.16 Gradio interface setup for BI-RADS classification demo

This segment defines the prediction function and the Gradio interface configuration (refer to figure 4.16). The predict function invokes the preprocessing step and feeds the image into the model, returning class probability percentages. Gradio's Interface constructs the web app UI, accepting image inputs and displaying the top predicted BI-RADS classes with confidence scores, making the system accessible and easy to use.

4.6 Conclusion

This chapter has elucidated the thorough implementation method of the mammography classification system, encompassing data preparation, augmentation, model training, evaluation, and deployment. The stringent preparation techniques successfully mitigated dataset issues including class imbalance and image variability, establishing a solid groundwork for deep learning. The comparative investigation of ResNet50, DenseNet121, and InceptionV3 illustrated the efficacy of transfer learning for precise BI-RADS classification, with each model exhibiting strengths in identifying varied mammographic characteristics.

The effective implementation of the DenseNet121 model on the Hugging Face platform demonstrates the research's practical utility, facilitating accessible and scalable utilisation by clinicians and researchers. This implementation phase connects theoretical model creation with practical application, facilitating subsequent improvements and incorporation into clinical operations. Future efforts will concentrate on broadening dataset diversity and improving model interpretability to augment diagnostic reliability and acceptance.

Chapter 5: Result and Discussion

5.1 Introduction

This chapter provides a comprehensive evaluation of the three CNN models DenseNet121, InceptionV3, and ResNet50 designed for mammographic BI-RADS classification. The evaluation emphasises many performance parameters, including confusion matrices, classification metrics such as recall, F1-score, specificity, ROC-AUC curves, and training dynamics shown by accuracy and loss over epochs. Each indicator offers distinct insights into the models' strengths and weaknesses, together providing a comprehensive picture of their diagnostic skills. This chapter seeks to determine the most effective model architecture for precise and dependable mammography classification by assessing outcomes both individually and comparatively, which is essential for clinical decision-making and early breast cancer diagnosis.

The chapter starts with a comprehensive exposition of confusion matrices for each model, emphasising patterns of accurate and erroneous classifications. The next sections examine quantitative classification measures, such as macro-averaged recall and specificity, to further assess sensitivity and accuracy. ROC-AUC studies provide threshold-independent evaluations of model discrimination, while training accuracy and loss curves illustrate learning behaviour and possible overfitting concerns. The chapter concludes with a direct comparison of all three models, integrating the data to ascertain which model optimally balances accuracy, generalisation, and therapeutic usefulness. This comprehensive assessment aims to inform further research and practical implementation in automated mammography analysis.

5.2 Confusion Matrix

Table 5.1 Confusion Matrix for DenseNet121 model

True\Predicted	B1	B3	B4	B5
B1	186	0	0	0
B3	4	163	19	0
B4	1	39	146	0
B5	5	1	0	180

Table 5.1 indicates that the confusion matrix for DenseNet121 has high classification accuracy, especially for the BI-RADS 1 category, where all 186 samples were accurately classified, reflecting exceptional performance in detecting normal or benign cases. Misclassifications were noted in the borderline categories: BI-RADS 3 samples were sometimes misclassified as BI-RADS 1 (4 instances) and BI-RADS 4 (19 instances), while a significant proportion of BI-RADS 4 samples (39 instances) were projected as BI-RADS 3. This indicates that the model struggles to differentiate across neighbouring BI-RADS categories, which is anticipated given the nuanced radiological variations within these classifications. In the most severe category, BI-RADS 5, DenseNet121 accurately categorised most samples, with little misunderstanding about 5 samples predicted as BI-RADS 1 and 1 sample as BI-RADS 3. The findings demonstrate DenseNet121's proficiency in accurately identifying malignant cases, while suggesting potential for enhancement in differentiating intermediate lesion severity (see Table 5.1).

Table 5.2 Confusion Matrix for InceptionV3 model

True\Predicted	B1	B3	B4	B5
B1	186	0	0	0
B3	41	113	29	3
B4	12	23	151	0
B5	8	1	0	177

Table 5.2 shows the confusion matrix of the InceptionV3 model, indicating a more dispersed pattern of misclassification in contrast to DenseNet121. Although BI-RADS 1

achieved flawless classification for 186 samples, the model exhibited increased ambiguity in the BI-RADS 3 and 4 categories. Instances of BI-RADS 3 were misclassified in various ways, with 41 projected as BI-RADS 1, 29 as BI-RADS 4, and 3 as BI-RADS 5, highlighting challenges in precisely delineating this intermediate severity group. Likewise, BI-RADS 4 exhibited significant uncertainty, with 12, 23, and 151 samples categorised as BI-RADS 1, 3, and 4 accordingly, highlighting difficulties in consistently recognising this classification. BI-RADS 5 exhibited a limited number of misclassifications (8 categorised as BI-RADS 1 and 1 as BI-RADS 3); nonetheless, the model maintained robust discriminating capability for the malignant category. The proliferation of mistakes underscores InceptionV3's susceptibility to minor radiological discrepancies while indicating an elevated incidence of false negatives in clinically significant categories (see table 5.2).

Table 5.3 Confusion Matrix for ResNet50 model

True\Predicted	B1	B3	B4	B5
B1	165	18	0	3
B3	0	121	65	0
B4	0	132	54	0
B5	2	0	0	184

Figure 5.3 displays the confusion matrix for ResNet50, showing a clear pattern of classification performance. The BI-RADS 1 category encountered misclassification, with 18 samples erroneously classified as BI-RADS 3 and 3 as BI-RADS 5, suggesting a propensity of the model to overestimate abnormalities in benign instances. There was significant misclassification overlap between BI-RADS 3 and BI-RADS 4; specifically, 65 BI-RADS 3 samples were incorrectly categorised as BI-RADS 4, while 132 BI-RADS 4 samples were erroneously forecasted as BI-RADS 3. This bilateral perplexity indicates challenges in distinguishing these intermediate categories, along with the intrinsic ambiguity in mammography interpretation. Notwithstanding these challenges, ResNet50 exhibited strong

detection capabilities for the BI-RADS 5 category, accurately classifying 184 out of 186 samples with little confusion. The matrix illustrates ResNet50's greater caution in detecting malignancy, however it encounters difficulties in differentiating intermediate severity levels (see Figure 5.3).

5.3 Classification Metric

Table 5.4 Classification Metric for DenseNet121 models

Class\Metric	Recall	F1-Score	Specificity	Support
B1	1	0.9738	0.9821	186
B3	0.8763	0.8380	0.9283	186
B4	0.7896	0.8319	0.9659	186
B5	0.9677	0.9836	1	186
Accuracy			0.9073	744
Macro Avg	0.9073	0.9068	0.9691	744
Weighted Avg	0.9073	0.9068	0.9691	744

According to the categorisation data shown in Figure 5.4, DenseNet121 demonstrates robust performance across most BI-RADS categories. The total accuracy is 90.73%, indicating that the model accurately identified most of the samples. The recall for BI-RADS 1 is flawless at 1.0000, indicating that the model consistently recognises benign instances without overlooking any real negatives. For BI-RADS 5, indicative of very suspicious malignant patients, the recall is notably high at 0.9677, further underscoring the model's efficacy in identifying severe pathology. The specificity values are notably remarkable, particularly a flawless specificity of 1.0000 for BI-RADS 5, indicating a minimal incidence of false positives for this essential category. The recall rates for BI-RADS 3 and BI-RADS 4 are somewhat lower, at 0.8763 and 0.7849 respectively, indicating the inherent challenge in differentiating these intermediate groups with overlapping radiological characteristics. The F1-scores closely

correlate with the recall patterns, exhibiting peak values at both benign and malignant extremes, while demonstrating somewhat lower values for intermediate categories, reflecting a balance of accuracy and sensitivity (see Figure 5.4).

Table 5.5 Classification Metric for InceptionV3 model

Class\Metric	Recall	F1-Score	Specificity	Support
B1	1	0.8591	0.8907	186
B3	0.6075	0.6997	0.9570	186
B4	0.8118	0.8251	0.9480	186
B5	0.9516	0.9672	0.9946	186
Accuracy			0.8427	744
Macro Avg	0.8427	0.8378	0.9476	744
Weighted Avg	0.8427	0.8378	0.9476	744

Figure 5.5 presents the classification report for the InceptionV3 model, with an overall accuracy of 84.27%. The model accurately identifies all BI-RADS 1 instances, achieving a recall of 1.0000, highlighting its dependable detection of benign mammograms. The memory performance for BI-RADS 3 declines, with a recall rate of 0.6075, indicating a significant number of misclassified BI-RADS 3 cases, potentially impacting clinical decision-making. BI-RADS 4 instances have a memory rate of 0.8118, but BI-RADS 5 has a superior remember rate of 0.9516. The specificity is robust across all categories, particularly for BI-RADS 5, which has a value of 0.9946, demonstrating the model's efficacy in reducing false positives. The F1-score outcomes reflect the recall pattern, exhibiting the greatest score in BI-RADS 5 and the lowest in BI-RADS 3, underscoring challenges in sensitivity for the borderline category (see Figure 5.5).

Table 5.6 Classification Metric for ResNet50

Class\Metric	Recall	F1-Score	Specificity	Support
B1	0.9892	0.9813	0.9910	186
B3	0.6183	0.5808	0.8297	186
B4	0.5	0.5314	0.8728	186
B5	0.9731	0.9864	1	186
Accuracy			0.7702	744
Macro Avg	0.7702	0.77	0.9234	744
Weighted Avg	0.7702	0.77	0.9234	744

The classification metrics for ResNet50, seen in Figure 5.6, indicate an overall accuracy of 77.02%, which is inferior to those of the other models. The recall for BI-RADS 1 is very high at 0.9892, indicating that the model effectively identifies benign instances. The recall for BI-RADS 3 and BI-RADS 4 is significantly reduced to 0.6183 and 0.5000, respectively, indicating that the model often misclassifies these intermediate categories. The recall for BI-RADS 5 is consistently high at 0.9731, which is essential for identifying malignant patients. The specificity values are consistently high for both benign and malignant groups, although diminished for the intermediate categories, indicating difficulties in preventing false negatives in these ambiguous classifications. The F1-score distribution mirrors that of memory, exhibiting worse performance for BI-RADS 3 and 4, which are clinically significant categories for more scrutiny (see Figure 5.6).

5.4 ROC-AUC Graph

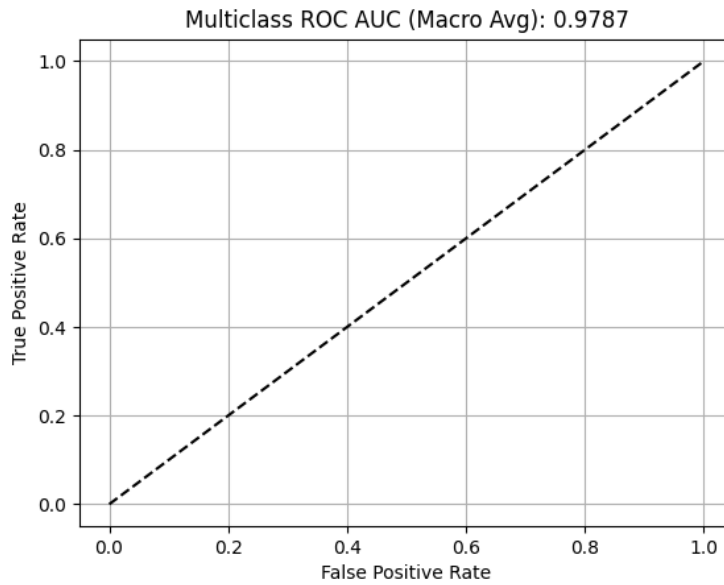


Figure 5.1 ROC-AUC Graph for DenseNet121 model

The ROC-AUC curve for DenseNet121, shown in Figure 5.1, demonstrates a substantial macro-averaged AUC value of 0.9787. This signifies that the model has exceptional discriminating ability across all BI-RADS categories. A ROC-AUC value around 1 indicates that DenseNet121 can proficiently differentiate between positive and negative occurrences across various threshold configurations, underscoring its efficacy in the classification job. The elevated AUC indicates the model's capacity to sustain an advantageous equilibrium between sensitivity and specificity, essential for clinical applications where both false negatives and false positives have substantial repercussions. The ROC-AUC curve substantiates DenseNet121's significant capability as a dependable diagnostic instrument for mammographic image categorization (see Figure 5.1).

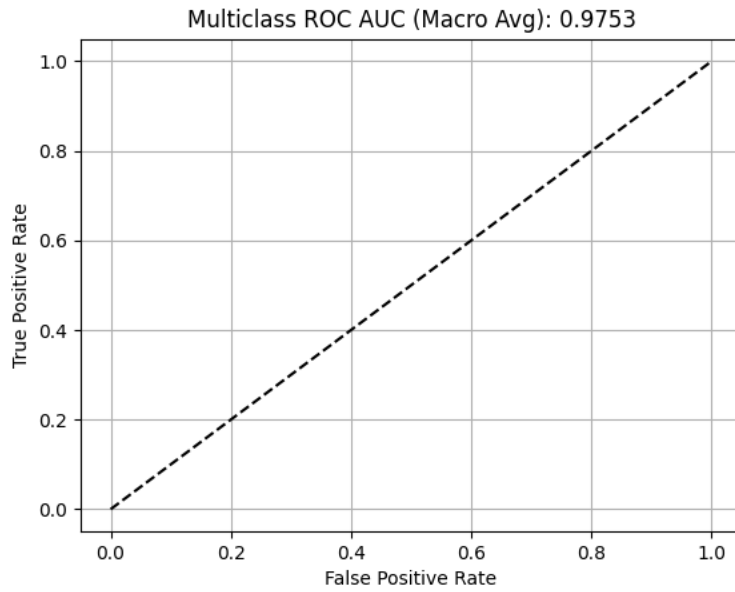


Figure 5.2 ROC-AUC Graph for Inceptionv3 model

Figure 5.2 illustrates the ROC-AUC curve for the InceptionV3 model. This model achieves a macro-averaged AUC of 0.9753. Although this value is slightly lower than that of DenseNet121, it nevertheless indicates that the classification performance remains commendable. According to the curve, InceptionV3 seems capable of effectively distinguishing among the numerous BI-RADS categories over a broad range of classification parameters. The model's use in clinical settings is validated by this level of performance, demonstrating its robustness in prediction accuracy. Nonetheless, the ROC-AUC, which is marginally worse than that of DenseNet121, may suggest a modest reduction in classification consistency across all categories (see Figure 5.2).

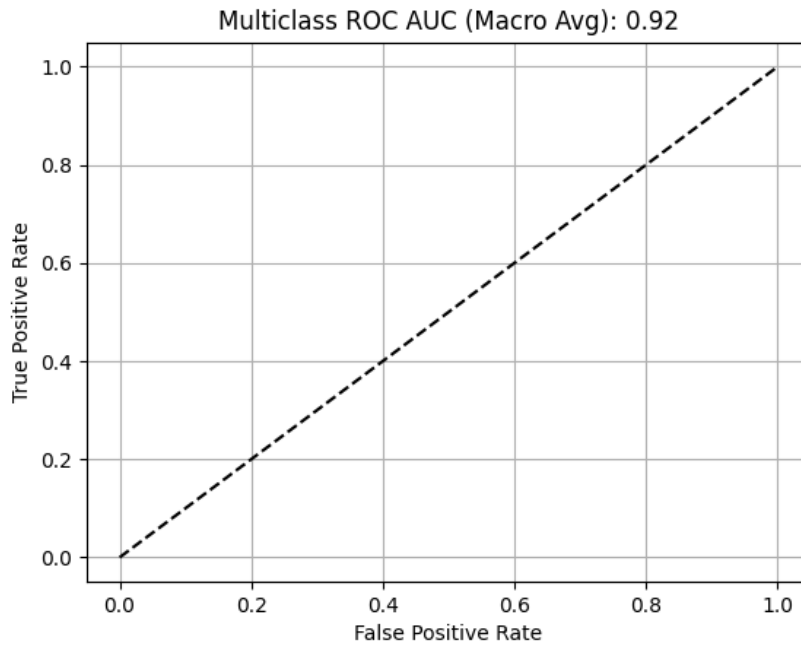


Figure 5.3 ROC-AUC Graph for ResNet50 model

In comparison to the other models, the ROC-AUC curve presented in Figure 5.3 for ResNet50 has a macro-averaged AUC of 0.92, signifying substantial discriminating power, but lower than that of the other models. Moreover, the difference indicates that ResNet50 may exhibit reduced sensitivity to variations in mammographic features throughout the BI-RADS spectrum. This is even though an AUC of 0.9 is generally considered high. This decline in performance may impact the model's reliability in detecting ambiguous or borderline events. Nonetheless, the ROC-AUC curve indicates that ResNet50 remains a feasible choice for breast cancer screening, while there is potential for enhancement in classification consistency (see Figure 5.3).

5.5 Accuracy and Loss over Epochs

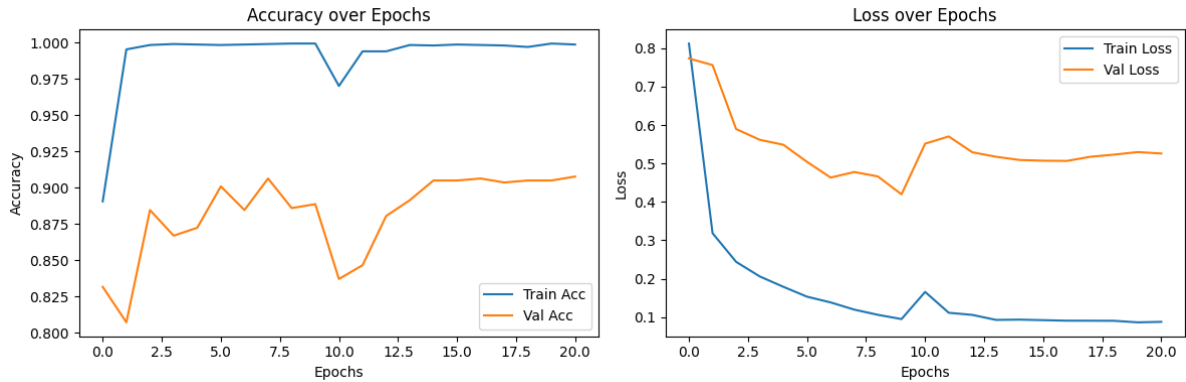


Figure 5.4 Accuracy and Loss over Epochs for DenseNet121 model

The accuracy and loss curves for DenseNet121 based on Figure 5.4 exhibit a pronounced initial learning phase, with training accuracy swiftly ascending to almost optimal values early in the training process, but validation accuracy increases more gradually, stabilizing at about 90%. The training loss consistently declines, signifying successful error reduction on the training dataset. The validation loss demonstrates swings and plateaus throughout the mid-training period, failing to improve beyond roughly epoch 10. The discrepancy between nearly flawless training accuracy and comparatively reduced, stabilized validation accuracy indicates slight overfitting. This overfitting is likely attributable to the model's complexity in relation to the constrained and potentially noisy mammography dataset, where subtle picture properties hinder generalization. Regularization methods, including L2 weight decay and dropout (rate 0.7), together with early halting with a patience of 4 epochs, seem to mitigate but not eradicate this overfitting. The observed training dynamics correspond with predictions for a deep convolutional architecture trained on medical pictures characterized by intrinsic unpredictability and a restricted sample size.

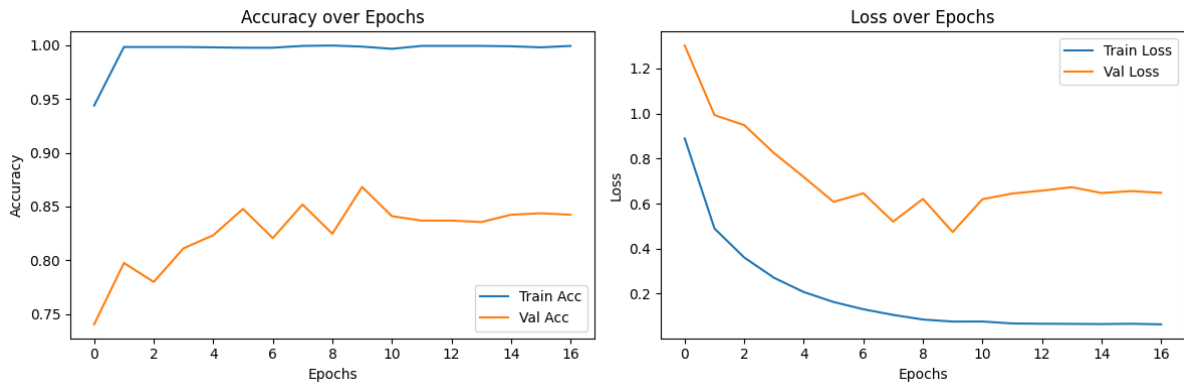


Figure 5.5 Accuracy and Loss over Epochs for InceptionV3 model

The InceptionV3 curves based on Figure 5.5 exhibit a same trend: training accuracy rapidly attains elevated levels near 100%, but validation accuracy trails, oscillating around 84%. The training loss decreases dramatically and continuously, but the validation loss exhibits fluctuations and stays comparatively elevated, suggesting difficulties in generalizing beyond the training dataset. The disparity between training and validation performance indicates overfitting, intensified by the model's complicated architecture and the dataset's intricacy. Dropout (rate 0.6) and L2 regularization were used, in conjunction with early stopping set at a patience of 6 epochs; nonetheless, the validation loss does not closely align with the training loss. The patterns indicate that although InceptionV3 excels at feature extraction, the quantity and diversity of training data constrain its generalization capabilities, a typical issue in specialized medical imaging applications with limited datasets.

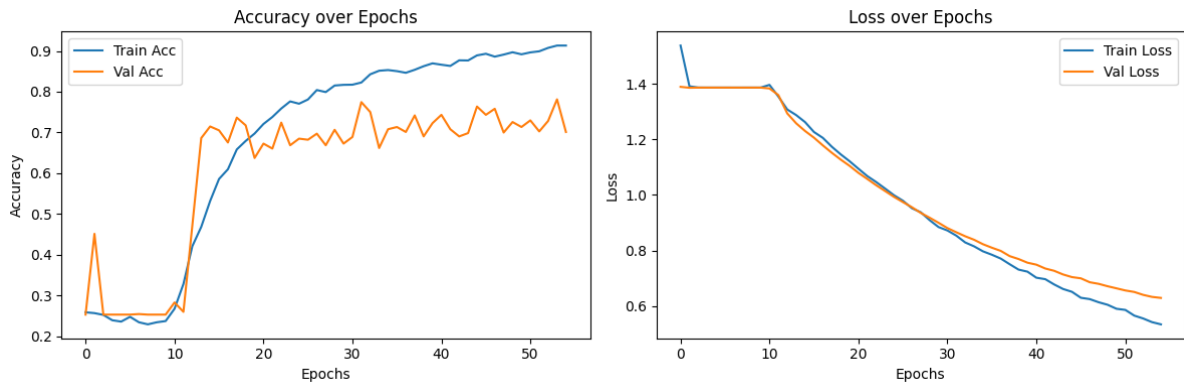


Figure 5.6 Accuracy and Loss over Epochs for Resnet50 model

The ResNet50 graphs based on Figure 5.6 display a significantly revised training profile. The training accuracy increases progressively, attaining almost 90% in the latter epochs, but the validation accuracy stabilizes around 70–75%, indicating a broader disparity compared to the other models. The training and validation loss curves initially align closely but subsequently diverge somewhat as training advances. This pattern suggests a blend of mild overfitting and potential early underfitting, as the model fails to adequately represent the data distribution. The lack of L2 regularization and a reduced dropout rate (0.5) relative to the other models may account for this result. Furthermore, implementing early stopping with a patience of 8 epochs extended the training period, thereby increasing the model's susceptibility to overfitting. In summary, ResNet50's training dynamics indicate that, although it learns consistently, it may not effectively balance bias and variance on this dataset compared to DenseNet121 and InceptionV3.

All models exhibit mild to moderate overfitting, mostly attributable to the dataset's short size and complexity, slight inter-class distinctions, and restricted data augmentation beyond rescaling. The intricacy of the models allows for the memorization of training instances; nonetheless, generalizing to unseen validation data continues to pose difficulties.

Regularization techniques and early halting alleviate, however do not completely rectify, this problem. Future endeavours may encompass enhanced augmentation, transfer learning from more extensive domain-specific datasets, or streamlined architectures to optimize the equilibrium between fitting and generalization.

5.6 Comparative Analysis of Models

Table 5.7 Details Analysis for all three models

Metric	DenseNet121	InceptionV3	Resnet50
Accuracy (%)	90.73	84.27	77.02
Macro Recall	0.9073	0.8427	0.7702
Macro F1-Score	0.9068	0.8378	0.77
Macro Specificity	0.9691	0.9476	0.9234
Macro Roc-AUC	0.9787	0.9753	0.92
Overfitting Level	Mild	Moderate	Moderate to High

Of the three assessed models based on Table 5.7, DenseNet121 substantially surpasses both InceptionV3 and ResNet50 on all significant metrics, establishing it as the optimal selection for the BI-RADS classification job. DenseNet121 attains the maximum accuracy at 90.73%, demonstrating its exceptional capability to accurately categorize mammography images. Its macro recall and F1-score over 0.90 indicate a constant ability to identify true positive cases while preserving an effective equilibrium between accuracy and sensitivity. The exceptionally high specificity (0.9691) signifies a reduced incidence of false positives, which is essential in medical diagnosis to prevent needless procedures.

Conversely, InceptionV3, albeit a formidable competitor, lags below DenseNet121 with an accuracy of 84.27% and much worse recall and F1-scores. This indicates that it is less dependable in finding positive instances, particularly in intermediate BI-RADS categories where categorization poses significant challenges. Despite its ROC-AUC being comparable to

that of DenseNet121, the other metrics indicate that its predictions are less reliable, perhaps attributable to mild overfitting and suboptimal feature extraction for this job.

ResNet50 has the lowest performance of the three, with a markedly diminished accuracy of 77.02% and decreased recall and F1-score, signifying its difficulty in accurately categorizing both benign and malignant pictures. Its increased propensity for overfitting and diminished specificity render it less appropriate for clinical use, where precision and dependability are critical.

DenseNet121 is distinguished as the most stable and dependable model because to its dense connection design, which facilitates superior feature propagation and reuse. The structural advantage of DenseNet121 enables it to discern more intricate patterns in mammographic pictures, leading to enhanced classification performance with little overfitting. Consequently, DenseNet121 is the optimal model for practical BI-RADS classification in clinical environments.

5.7 Conclusion

In summary, the comparison analysis reveals that DenseNet121 surpasses InceptionV3 and ResNet50 on several essential measures, confirming its status as the preeminent model for BI-RADS mammographic categorization. DenseNet121's elevated accuracy, balanced recall, good specificity, and superior ROC-AUC demonstrate its efficacy in accurately distinguishing between benign and malignant breast lesions with minimum misclassification. The architectural benefits, such as dense connectivity and efficient regularization, provide superior feature extraction and generalization, even in the face of obstacles presented by the intricate and constrained dataset. Although InceptionV3 demonstrates competitive performance, especially in ROC-AUC, its diminished recall and increased danger of overfitting undermine its dependability. ResNet50's somewhat worse performance and heightened vulnerability to

overfitting indicate it is less appropriate for this task. This chapter's findings highlight DenseNet121 as the optimal deep learning model for mammography classification, suggesting significant potential for improving breast cancer screening accuracy and clinical processes.

Chapter 6: Conclusion and Future Work

6.1 Introduction

This chapter encapsulates the study's results, addressing the contributions, limitations, and prospective future avenues for research. The primary aim of the research was to assess the efficacy of deep learning models like ResNet, DenseNet, and InceptionV3 in categorizing mammography images based on the BI-RADS system. The models were trained and tested with a dataset from Kaggle, with their performance examined by measures like accuracy, recall, specificity, F1-score, and AUC-ROC. The study illustrates the capability of deep learning to improve breast cancer detection by automating the BI-RADS classification, hence facilitating clinical workflows and aiding radiologists in achieving more precise and efficient diagnoses.

6.2 Contribution

This paper provides a substantial technological addition to medical image analysis by performing a comparative assessment of three deep learning architectures: ResNet50, DenseNet121, and InceptionV3, for the automatic categorisation of mammography images based on BI-RADS categories. Each model was executed using a two-phase transfer learning methodology, whereby pre-trained weights on ImageNet were refined for the domain-specific goal of BI-RADS classification. The models' performance was meticulously evaluated using clinically relevant measures such as accuracy, recall, specificity, F1-score, and AUC-ROC, offering a thorough insight into the diagnostic efficacy of each architecture. DenseNet121 proved to be the most successful model, with a macro-averaged AUC of 0.9787 and an accuracy of 90.73%, indicating robust classification performance across all designated BI-RADS levels (1, 3, 4, and 5).

An essential technical insight derived from the research is the examination of misclassification patterns, especially within the confusion matrices of each model. The results

indicated that most classification mistakes transpired between BI-RADS 3 and 4, highlighting the clinical uncertainty often associated with intermediate lesion categories. DenseNet121 shown enhanced specificity in differentiating benign from malignant instances, while ResNet50, despite its efficacy in identifying high-suspicion lesions (BI-RADS 5), had difficulties with intermediate classifications. This comparative investigation underscores the stability and generalisation capabilities of DenseNet121 while establishing a basis for future enhancements in feature interpretability and decision support systems in deep learning-based diagnostics.

The work highlights its practical contribution by successfully deploying DenseNet121 on a cloud-based platform using Gradio and Hugging Face, in addition to performance assessment. This implementation converts the learnt model into an accessible, interactive online service enabling physicians and researchers to input mammographic pictures and get real-time BI-RADS classifications with associated confidence ratings. The deployment demonstrates the practicality of incorporating artificial intelligence into clinical workflows, providing a scalable and user-friendly solution to aid radiologists in enhancing diagnostic consistency and early breast cancer detection.

6.3 Limitation

The main limitation of this research is the significant class imbalance in the dataset, especially with BI-RADS 5, which has much less photos than other categories. Despite the use of data augmentation to equilibrate the classes and provide sufficient representation for underrepresented categories, the intrinsic imbalance may still affect the model's efficacy. This imbalance may result in models that excel at identifying prevalent categories (such as BI-RADS 1) but falter in properly classifying more nuanced or intermediate instances like BI-RADS 3 and 4, which exhibit comparable characteristics in mammographic pictures.

Another limitation of the research is the quantity of the dataset and the resolution at which the photos were analysed. This research used a resolution of 327x541 pixels to maintain anatomical details and avoid distortion during preprocessing, in contrast to the common practice of using lesser input sizes such as 224x224 pixels. The elevated resolution guarantees the collection of intricate information in mammograms; nevertheless, it also escalates computing demands, possibly hindering the speed of model training and assessment. The dataset's size, while considerable, may not adequately capture the variety of mammographic pictures present in actual clinical settings, hence limiting the model's generalizability across various populations or imaging systems.

The research concentrated on BI-RADS levels 1, 3, 4, and 5, excluding levels 0, 2, and 6. BI-RADS 0 and 2 signify categories that do not need further evaluation (BI-RADS 2) or require supplementary imaging (BI-RADS 0), while BI-RADS 6 indicates verified cancer cases. By excluding these levels, the research focused on categories with differing levels of cancer suspicion; nonetheless, this limitation hindered the model's capacity to learn from the whole spectrum of mammographic categories. Incorporating all BI-RADS levels, from 0 to 6, would have offered a more thorough perspective on the model's classification skills and more accurately represented the intricacies of breast cancer detection in real practice.

6.4 Future works

A significant approach for future study is the augmentation of the dataset to include a broader and more varied assortment of mammograms. The present research used the Kaggle dataset, which, albeit containing useful information, is restricted in size and variety. Future endeavors should focus on integrating supplementary datasets from diverse healthcare facilities, including photos from various geographic locations, clinical environments, and demographic

cohorts. This variety would enhance the generalizability of the models, assuring their efficacy across various patient groups and imaging modalities.

Furthermore, including datasets with diverse picture quality and multiple mammographic perspectives might enhance the model's robustness and reliability in practical clinical settings. Additional investigation into ensemble techniques is a possible avenue for future research. This work concentrated on singular deep learning models; however, integrating the advantages of numerous models might markedly enhance prediction efficacy. Employing strategies like as bagging, boosting, or stacking may alleviate the limitations of individual models, resulting in enhanced generalization and improved classification accuracy. An ensemble of ResNet, DenseNet, and InceptionV3 models may outperform individual models, particularly when trained on varied datasets. Ensemble approaches may be effective in reducing overfitting and enhancing model stability when confronted with novel or noisy data.

A promising avenue for future research is the incorporation of the model into actual clinical operations. The optimal model (DenseNet121) has been implemented on a cloud-based platform for demonstration purposes; however, further research should concentrate on its incorporation into real clinical settings. The implementation of real-time use, allowing radiologists to input mammography pictures and get instantaneous BI-RADS classifications, might significantly improve the efficiency of breast cancer screening. Moreover, it is important to emphasize the model's scalability and performance in managing extensive clinical data, guaranteeing its capacity to analyse pictures swiftly and precisely in a high-throughput environment. This integration will be essential for converting AI-driven diagnostic tools from research prototypes to practical, widely utilized clinical applications.

6.5 Conclusion

In summary, this research underlines the potential of deep learning models in the automated classification of mammography, with a particular emphasis on BI-RADS levels 1, 3, 4, and 5. The assessment of three distinct models like ResNet, DenseNet, and InceptionV3 revealed that DenseNet121 above the others on accuracy, recall, specificity, and AUC-ROC. Despite constraints like class imbalance and a relatively small dataset, the results highlight the potential of AI to improve diagnosis accuracy in breast cancer screening. Subsequent efforts should prioritize the augmentation of the dataset, the inclusion of clinical attributes, and the assimilation of the models into practical clinical settings to enhance their dependability and scalability.

References

- Ashwini, P., Suguna, N., & Vadivelan, N. (2024). Detection and classification of breast cancer types using VGG16 and ResNet50 deep learning techniques. *International Journal of Electrical and Computer Engineering*, 14(5), 5481–5488. <https://doi.org/10.11591/ijece.v14i5.pp5481-5488>
- Bhushan, A., Gonsalves, A., & Menon, J. U. (2021). Current state of breast cancer diagnosis, treatment, and theranostics. In *Pharmaceutics* (Vol. 13, Issue 5). MDPI. <https://doi.org/10.3390/pharmaceutics13050723>
- Daniel López-Cabrera, J., Alberto López Rodríguez, L., & Pérez-Díaz, M. (2020). Classification of breast cancer from digital mammography using deep learning. *Inteligencia Artificial*, 23(65), 56–66. <https://doi.org/10.4114/intartif.vol23iss65pp56-66>
- Gandomkar, Z., Suleiman, M. E., Demchig, D., Brennan, P. C., & McEntee, M. F. (2019). *BI-RADS density categorization using deep neural networks*. 22. <https://doi.org/10.1117/12.2513185>
- Gao, Y., Lin, J., Zhou, Y., & Lin, R. (2023). The application of traditional machine learning and deep learning techniques in mammography: a review. In *Frontiers in Oncology* (Vol. 13). Frontiers Media SA. <https://doi.org/10.3389/fonc.2023.1213045>
- Ginsburg, O., Yip, C. H., Brooks, A., Cabanes, A., Caleffi, M., Yataco, J. A. D., Gyawali, B., McCormack, V., de Anderson, M. M. L., Mehrotra, R., Mohar, A., Murillo, R., Pace, L. E., Paskett, E. D., Romanoff, A., Rositch, A. F., Scheel, J. R., Schneidman, M., Unger-Saldaña, K., ... Anderson, B. O. (2020). Breast Cancer Early Detection: A Phased Approach to Implementation. *Cancer*, 126(S10), 2379–2393. <https://doi.org/10.1002/cncr.32887>
- Khanna, P., Sahu, M., Singh, B. K., & Bhateja, V. (2022). Combining modified hyper learning binary dragonfly algorithm and deep learning for BI-RADS classification of breast masses in mammograms. *Expert Systems*. <https://doi.org/10.1111/exsy.13200>
- Kim, S. T., Lee, J. H., Lee, H., & Ro, Y. M. (2018). Visually interpretable deep network for diagnosis of breast masses on mammograms. *Physics in Medicine and Biology*, 63(23). <https://doi.org/10.1088/1361-6560/aaef0a>
- Liu, H., Chen, Y., Zhang, Y., Wang, L., Luo, R., Wu, H., Wu, C., Zhang, H., Tan, W., Yin, H., & Wang, D. (2021). A deep learning model integrating mammography and clinical factors facilitates the malignancy prediction of BI-RADS 4 microcalcifications in breast cancer screening. *European Radiology*, 31(8), 5902–5912. <https://doi.org/10.1007/s00330-020-07659-y>
- Liu, Z., Peng, J., Guo, X., Chen, S., & Liu, L. (2024). Breast cancer classification method based on improved VGG16 using mammography images. *Journal of Radiation Research and Applied Sciences*, 17(2), 100885. <https://doi.org/10.1016/j.jrras.2024.100885>
- Macit, H. S., & Sabanci, K. (2023). Benchmarking of ResNet models for breast cancer diagnosis using mammographic images. *International Journal of Applied Methods in Electronics and Computers*. <https://doi.org/10.58190/ijamec.2023.39>

- Pawar, S. D., Sharma, K. K., Sapate, S. G., Yadav, G. Y., Alroobaea, R., Alzahrani, S. M., & Hedabou, M. (2022). Multichannel DenseNet Architecture for Classification of Mammographic Breast Density for Breast Cancer Detection. *Frontiers in Public Health*, 10. <https://doi.org/10.3389/fpubh.2022.885212>
- Pullaiah, N. R. P., Venkatesekhar, D., Venkatramana, P., & Sudhakar, B. (2022). Breast Cancer Classification Using Customized ResNet Based Convolution Neural Networks. *International Journal of Intelligent Engineering and Systems*, 15(6), 382–391. <https://doi.org/10.22266/ijies2022.1231.35>
- Rahman, H., Naik Bukht, T. F., Ahmad, R., Almadhor, A., & Javed, A. R. (2023). Efficient Breast Cancer Diagnosis from Complex Mammographic Images Using Deep Convolutional Neural Network. *Computational Intelligence and Neuroscience*, 2023(1). <https://doi.org/10.1155/2023/7717712>
- Ramadan, S. Z. (2020). Methods Used in Computer-Aided Diagnosis for Breast Cancer Detection Using Mammograms: A Review. *Journal of Healthcare Engineering*, 2020. <https://doi.org/10.1155/2020/9162464>
- Ramadhani, S., Islam, U., Sultan, N., Riau, S. K., Soebrantas, H. R., Panam, R., 15, K., Madani, T., Tampan, K., & Kampar, K. (2021). A Review Comparative Mammography Image Analysis on Modified CNN Deep Learning Method. *Indonesian Journal of Artificial Intelligence and Data Mining (IJAIMD)*, 4(1), 54–61. <https://doi.org/10.24014/ijaidm.v4i1.10891>
- Roth, H. R., Chang, K., Singh, P., Neumark, N., Li, W., Gupta, V., Gupta, S., Qu, L., Ihsani, A., Bizzo, B. C., Wen, Y., Buch, V., Shah, M., Kitamura, F., Mendonça, M., Lavor, V., Harouni, A., Compas, C., Tetreault, J., ... Kalpathy-Cramer, J. (2020). *Federated Learning for Breast Density Classification: A Real-World Implementation*. https://doi.org/10.1007/978-3-030-60548-3_18
- Sabani, A., Landsmann, A., Hejduk, P., Schmidt, C., Marcon, M., Borkowski, K., Rossi, C., Ciritsis, A., & Boss, A. (2022). BI-RADS-Based Classification of Mammographic Soft Tissue Opacities Using a Deep Convolutional Neural Network. *Diagnostics*, 12(7). <https://doi.org/10.3390/diagnostics12071564>
- Sadad, T., Khan, A. R., Hussain, A., Tariq, U., Fati, S. M., Bahaj, S. A., & Munir, A. (2021). Internet of medical things embedding deep learning with data augmentation for mammogram density classification. *Microscopy Research and Technique*, 84(9), 2186–2194. <https://doi.org/10.1002/jemt.23773>
- Shi, P., Wu, C., Zhong, J., & Wang, H. (2019). Deep learning from small dataset for bi-rads density classification of mammography images. *Proceedings - 10th International Conference on Information Technology in Medicine and Education, ITME 2019*, 102–109. <https://doi.org/10.1109/ITME.2019.00034>
- Trayes, K. P., & Cokenakes, S. E. (2021). Breast Cancer Treatment. *American Family Physician*, 2, 171–178.
- Wang, L. (2024). Mammography with deep learning for breast cancer detection. In *Frontiers in Oncology* (Vol. 14). Frontiers Media SA. <https://doi.org/10.3389/fonc.2024.1281922>

Watkins, E. J. (2019). Overview of breast cancer. *Journal of the American Academy of Physician Assistants*, 32(10), 13–17. <https://doi.org/10.1097/01.JAA.0000580524.95733.3d>

Zhang, E., & Gu, X. (2020). *BIRADS Features-Oriented Semi-supervised Deep Learning for Breast Ultrasound Computer-Aided Diagnosis*. <https://doi.org/10.48550/arXiv.1904.01076>