Impact of Advanced Feature Engineering on Machine Learning Models in Multimodal Breast Cancer Diagnosis

by

Ene Joyce Ojaide

in

2024

ABSTRACT

This study addresses the challenge of improving breast cancer diagnosis accuracy through advanced machine learning techniques applied to multimodal imaging data. Despite recent progress, limitations in accuracy and interpretability persist, particularly when integrating multiple imaging modalities. This research combines sophisticated feature engineering with deep learning models to enhance diagnostic precision and clinical relevance.

The study utilises BreakHis for histopathological images and CBIS-DDSM for mammographic images, enabling comprehensive analysis at cellular and anatomical levels. A multi-stage feature engineering pipeline incorporates texture analysis, shape descriptors, and colour features. The research compares machine learning models in single-modality and multimodal settings, including traditional algorithms (SVM, Random Forest) and deep learning architectures (Custom CNN, ResNet50, DenseNet121).

Findings demonstrate the significant impact of feature engineering, with the SVM model showing a 25.26% improvement in the Matthews Correlation Coefficient. The dual-stream DenseNet121 model, integrating both imaging modalities, achieves 91.92% accuracy and 0.9699 ROC AUC, surpassing single-modality approaches. The study addresses model interpretability through feature importance analysis, which is vital for clinical trust and adoption.

This research provides insights into effective multimodal integration strategies and synergistic effects of feature engineering with deep learning in medical imaging. The findings have implications for developing accurate, interpretable, and clinically applicable AI-assisted diagnostic tools for breast cancer, potentially improving early detection rates and patient outcomes.

TABLE OF CONTENTS

ABSTRACT 2

ACKNOWLEDGEMENTS .... ERROR! BOOKMARK NOT DEFINED.

TABLE OF CONTENTS 3

LIST OF FIGURES 8

LIST OF TABLES 11

CHAPTER 1 12

INTRODUCTION 12

INTRODUCTION 12

PROBLEM STATEMENT 13

AIMS AND OBJECTIVES 14

SCOPE OF THE STUDY 15

SIGNIFICANCE OF THE STUDY 16

TIME PLAN 17

CHAPTER 2 2

LITERATURE REVIEW 2

INTRODUCTION 2

SOURCES 3

CRITICAL EVALUATION 6

TRADITIONAL MACHINE LEARNING ALGORITHMS: 6

DEEP LEARNING TECHNIQUES: 9

MULTIMODAL APPROACHES AND DATA FUSION 12

Transfer Learning and Pre-trained Models 14

Performance Comparison and Review Studies 15

RESEARCH GAPS AND QUESTIONS 17

CHAPTER 3 24

NEW IDEAS & APPROACH 24

INTRODUCTION 24

METHODOLOGY 24

Data Acquisition 25

Hardware Specifications and Computational Resources 26

Exploratory Data Analysis (EDA) 26

Data Preprocessing 28

Data Partitioning 29

MULTIMODAL FEATURE ENGINEERING FRAMEWORK 30

MODEL DEVELOPMENT AND ARCHITECTURE 33

MODEL TRAINING AND OPTIMISATION 35

MODEL EVALUATION METRICS 36

MODEL VISUALISATION AND INTERPRETATION 37

COMPARATIVE ANALYSIS WITH AND WITHOUT FEATURE ENGINEERING 38

Impact of Feature Engineering 38

Single Modality vs. Multimodal Performance 38

Effect of Model Complexity 38

Incorporation of Engineered Features in Deep Learning Models 39

Clinical Relevance 39

PROFESSIONAL, SOCIAL, ETHICAL, AND LEGAL CONSIDERATIONS 39

PROFESSIONAL CONSIDERATIONS 39

SOCIAL CONSIDERATIONS... 40

ETHICAL CONSIDERATIONS... 40

LEGAL CONSIDERATIONS 40

DATA HANDLING AND REPRODUCIBILITY 40

BIAS AND GENERALISABILITY 41

CHAPTER 4 42

IMPLEMENTATION 42

INTRODUCTION 42

DATA ACQUISITION AND PREPROCESSING OF BREAKHIS AND CBIS-DDSM DATASETS 42

EXPLORATORY DATA ANALYSIS... 46

MAGNIFICATION LEVEL ANALYSIS (BREAKHIS)... 47

TEXTURE ANALYSIS FOR BREAKHIS DATASET 48

COLOUR ANALYSIS (BREAKHIS) 49

METADATA PROCESSING AND ANALYSIS (CBIS-DDSM)....**ERROR! BOOKMARK NOT DEFINED.**

LESION TYPE AND BREAST DENSITY ANALYSIS ....**ERROR! BOOKMARK NOT DEFINED.**

IMAGE QUALITY ASSESSMENT (CBIS-DDSM).. 50

DIMENSIONALITY REDUCTION AND VISUALISATION OF DATASETS......................... 51

STATISTICAL ANALYSIS 52

DATA PREPARATION, PREPROCESSING AND FEATURE EXTRACTION 54

MODEL IMPLEMENTATION 58

MODEL EVALUATION AND VISUALISATION 64

CHAPTER 5 64

EVALUATION.. 66

INTRODUCTION **66**

OVERVIEW OF EVALUATION METHODOLOGY .... ERROR! BOOKMARK NOT DEFINED.

EVALUATION METRICS AND STATISTICAL METHODS **67**

PERFORMANCE METRICS 67

STATISTICAL TESTS AND EFFECT SIZE CALCULATIONS 67

SINGLE MODAL APPROACHES **68**

TRADITIONAL MACHINE LEARNING WITHOUT FEATURE ENGINEERING 68

TRADITIONAL MACHINE LEARNING WITH FEATURE ENGINEERING 71

MULTIMODAL APPROACHES **74**

DUAL-STREAM MODELS WITH AND WITHOUT ADDITIONAL FEATURE ENGINEERING 74

COMPARATIVE ANALYSIS 76

IMPACT OF FEATURE ENGINEERING 80

CONTEXTUALISATION OF RESULTS 84

DISCUSSION OF FINDINGS **85**

LIMITATIONS AND SUCCESSES **87**

CRITICAL ASSESSMENT OF STUDY LIMITATIONS 87

THE SUCCESS OF THE STUDY 88

CHAPTER 6 .... **90**

CONCLUSIONS / FUTURE WORK **90**

INTRODUCTION **90**

SUMMARY OF KEY FINDINGS **90**

PERFORMANCE OF MACHINE LEARNING MODELS 90

EFFECTIVENESS OF MULTIMODAL INTEGRATION 91

MODEL INTERPRETABILITY AND CLINICAL RELEVANCE 91

IMPLICATIONS FOR BREAST CANCER DIAGNOSIS 92

FUTURE RESEARCH DIRECTIONS 93

CONCLUSION 93

REFERENCES 95

APPENDIX A 103

APPENDIX B 104

APPENDIX C 105

APPENDIX D 106

APPENDIX E 107

APPENDIX F 108

LIST OF FIGURES

Figure 1: Gantt Chart 1

Figure 2: *Machine Learning Overview.* 2

Figure 3: Methodology Flowchart 25

Figure 4: EDA Workflow... 26

Figure 5: Data Preprocessing Workflow.. 28

Figure 6: Feature Engineering 31

Figure 7: Model Architecture Comparison.. 33

Figure 8: Training and Optimization Flowchart... 35

Figure 9: Evaluation Metrics.. 36

Figure 10: Comparative Analysis Framework 38

Figure 11:Data Preprocessing and Checkpoint Creation Workflow 45

Figure 12: Sample Images from BreakHis Dataset 46

Figure 13: Sample Images from CBIS-DDSM Dataset.................................. 47

Figure 14: Class Distribution Across Magnification Levels in BreakHis Dataset . 47

Figure 15: Texture Analysis for Images from the BreakHis Dataset 48

Figure 18: Distribution of Breast Density Categories In CBIS-DDSM Dataset ... 50

Figure 19: Image Quality Metrics Distribution in CBIS-DDSM Dataset............. 51

Figure 20: t-SNE and UMAP Visualizations of BreakHis................................. 52

Figure 22: Correlation Heatmap of Key Features in the CBIS-DDSM Dataset ... 53

Figure 23: Custom CNN Architecture Diagram 59

Figure 24: Dual-Stream CNN Architecture Diagram 61

Figure 25: Dual-Stream ResNet50 Architecture Diagram.............................. 62

Figure 27: Graphical representation of Model Evaluation 69

Figure 28: Random Forest features 71

Figure 29: Single Modality Deep Learning Evaluation................................... 73

Figure 30: Evaluation Metrics for Dual-Stream Models 75

Figure 31: Statistical Tests.. 80

Figure 32: Impact of Feature Engineering 82

Appendix 1:Confusion Matrix and ROC curves for SVM Model without Feature Engineering 103

Appendix 2:Confusion Matrix and ROC Curves for Random Forest Model 103

Appendix 4: Confusion Matrix and ROC Curves for Random Forest Model with Feature Engineering. 104

Appendix 6: Confusion Matrix and ROC Curve for Single Modal Custom CNN . 104

Appendix 7: Training and validation Accuracy/Loss Curve for single modal Resnet 50 105

Appendix 8: Confusion Matrix and ROC Curve for single modal Resnet50 105

Appendix 10: Training and Validation Accuracy/Loss Curves for Dual-Stream Custom CNN 106

Appendix 11: Confusion Matrix and ROC Curve for Dual-Stream Custom CNN 106

Appendix 12: Training and Validation Accuracy/Loss Curves for Dual-Stream ResNet50 106

Appendix 14: Confusion Matrices and ROC Curves for Dual-Stream DenseNet121 107

Appendix 15: Training and Validation Accuracy/Loss Curves for dual stream custom CNN with Feature Engineering 108

Appendix 16: Confusion Matrix and ROC Curve for dual stream custom CNN with Feature Engineering 108

LIST OF TABLES

Table 1: Summary of Reviewed Literature.. 21

Table 2: Class Distribution Across Training, Validation, and Test Sets 54

Table 3: Evaluation of single modality Models without feature Engineering...... 69

Table 4: The top five features of Random Forest.. 70

Table 5: Evaluation of single modality models with Feature Engineering 71

Table 6: SVN and Random Forest Gain with Feature Engineering 72

Table 7: Evaluation of Custom CNN, Resnet50, and Densenet121 on single modality 73

Table 8: Evaluation of Dual Stream Models 74

Table 9: Summary of all model performances... 76

Table 10: Table of Statistical Test Results 79

Table 11: Impact of feature engineering on models..................................... 81

Table 12: Comparison of best results with published benchmarks 84

CHAPTER 1

INTRODUCTION

Introduction

Breast cancer continues to be a significant global public health issue affecting individuals of all genders. Early identification is vital in enhancing patient outcomes. Although advancements have been achieved in medical technologies, current diagnostic processes meet several obstacles, such as concerns about availability, accuracy, and precision. According to Cancer Research UK (n.d.), breast cancer caused around 685,000 deaths worldwide in 2020, and there were roughly 18.1 million new instances of cancer recorded globally. According to the World Health Organisation (WHO, 2022), there is a projected rise to 2.7 million new breast cancer cases by 2024.

The disease has various manifestations; the diagnostic process is further complicated as there is no single definitive method of detection (Eftekharian, et al. 2023). The current diagnostic paradigm, while practical, often results in prolonged processes due to the need for multiple tests. This challenge, coupled with the limited efficacy of existing technologies in managing advanced-stage patients, underscores the critical need for improved diagnostic tools and methodologies. Machine learning and artificial intelligence advancements have shown promising outcomes in enhancing diagnostic precision and effectiveness across various medical domains. However, breast cancer diagnosis presents unique challenges that require innovative approaches, particularly in extracting relevant information from complex medical images and constructing interpretable models that can gain the trust of medical experts (Das, et al. 2024).

Machine learning and artificial intelligence advancements have proved encouraging outcomes in improving diagnostic precision and effectiveness in diverse medical domains. However, when it comes to diagnosing breast cancer, specific problems and opportunities need to be considered. The extraction of relevant information from complicated medical pictures and the construction of interpretable models that can get the trust of medical experts are innovative approaches necessary for integration into other imaging modalities.

Problem Statement

Despite progress in breast cancer diagnosis, existing methods face limitations. Different imaging techniques render tissues inconsistently, hindering the development of uniform diagnostic criteria. Subtle distinctions between benign and malignant lesions often lead to misdiagnoses, necessitating advanced analysis methods. Limited integration of complementary data from multiple imaging modalities restricts holistic diagnostic views. Practical feature engineering and fusion approaches are needed to utilise diverse data sources, capitalising on each imaging modality's strengths.

The research gap in thoroughly studying feature engineering methods with supervised ML and DL algorithms for breast cancer classification is paramount. While some studies address feature selection and extraction, the literature lacks a discussion of engineering features to enhance model generalisability, accuracy, and clinical acceptability. This gap underscores the urgency and significance of the work being done in this field, as it directly impacts the accuracy and effectiveness of breast cancer diagnosis and treatment.

These challenges result in missed diagnoses, false positives, and treatment delays, underscoring the urgent need for more accurate, efficient, and interpretable diagnostic tools. An ideal solution would integrate multi-modal imaging information to improve early detection and classification of breast cancer while providing precise, explainable results to support clinical decision-making.

Aims and Objectives

Aims

This project aims to use advanced supervised machine learning techniques to create a dependable and innovative model for breast cancer diagnosis using multimodal imaging data. It seeks to apply innovative methods and feature engineering to enhance classification performance, thereby contributing to earlier detection and improved patient survival rates (Gray 2021).

Objectives

The objectives of the project are as follows:

- To critically evaluate recent advancements in machine learning and deep learning techniques for diagnosing breast cancer, with a focus on advanced feature engineering and multimodal integration techniques.
- To clean, process, and perform exploratory data analysis on the BreakHis and CBIS-DDSM datasets, ensuring data quality and consistency for later analysis.
- To develop and evaluate advanced feature engineering techniques and examine their effects on the classification performance of selected models.
- To develop and implement a custom neural network architecture that processes dual input streams from histopathological and mammographic images and evaluate and compare the performance of models with and without feature engineering.
- To analyse the results, discuss the findings, and suggest improvements for multimodal feature engineering in breast cancer diagnosis.

The objectives of this research project align with the SMART criteria as they are

Specific: Each objective targets a distinct aspect of the research process.

Measurable: Performance metrics and evaluation criteria are clearly defined.

Achievable: The goals are realistic within the project's scope and available resources.

Relevant: All objectives contribute directly to the overarching aim of improving breast cancer diagnosis.

Time-bound: The project timeline (as detailed in the Time Plan section) provides clear deadlines.

Scope of the Study

This study goes beyond simple single-modality analysis; it suggests a multimodal approach to diagnosing breast cancer that combines histopathological images from the BreakHis dataset with mammographic images from the CBIS-DDSM dataset. This will create a unique combination of cellular-level and anatomical information. This integration presents a significant challenge, requiring methods to effectively combine and analyse data from two distinct imaging modalities. The project will further explore advanced deep learning architectures capable of processing dual input streams simultaneously, pushing the boundaries of current machine learning applications in medical imaging. This includes developing custom neural network architectures for multimodal breast cancer image data complexities.

A vital aspect of this research is the exploration of feature engineering techniques tailored to each imaging modality. This includes implementing advanced texture analysis methods for histopathological images and sophisticated region-of-interest extraction techniques for mammograms. The study will use relevant interpretability techniques to enhance model transparency and trustworthiness. The scope extends to a systematic evaluation of various feature engineering methods, encompassing feature selection, dimensionality reduction (e.g., principal component analysis, linear discriminant analysis), feature scaling, and feature construction (Zouhri, et al. 2024; Al-Qazzaz, et al. 2023). This holistic approach will advance our understanding of the underlying data characteristics and pave the way for more robust, interpretable, and clinically applicable ML and DL models. Moreover, given the requirement for a more powerful GPU, Google Colab will be employed for model development and training, using its GPU capabilities for efficient computation. This platform choice ensures the research's reproducibility and accessibility while demonstrating the feasibility of implementing complex machine-learning models using widely available cloud-based resources.

Significance of the Study

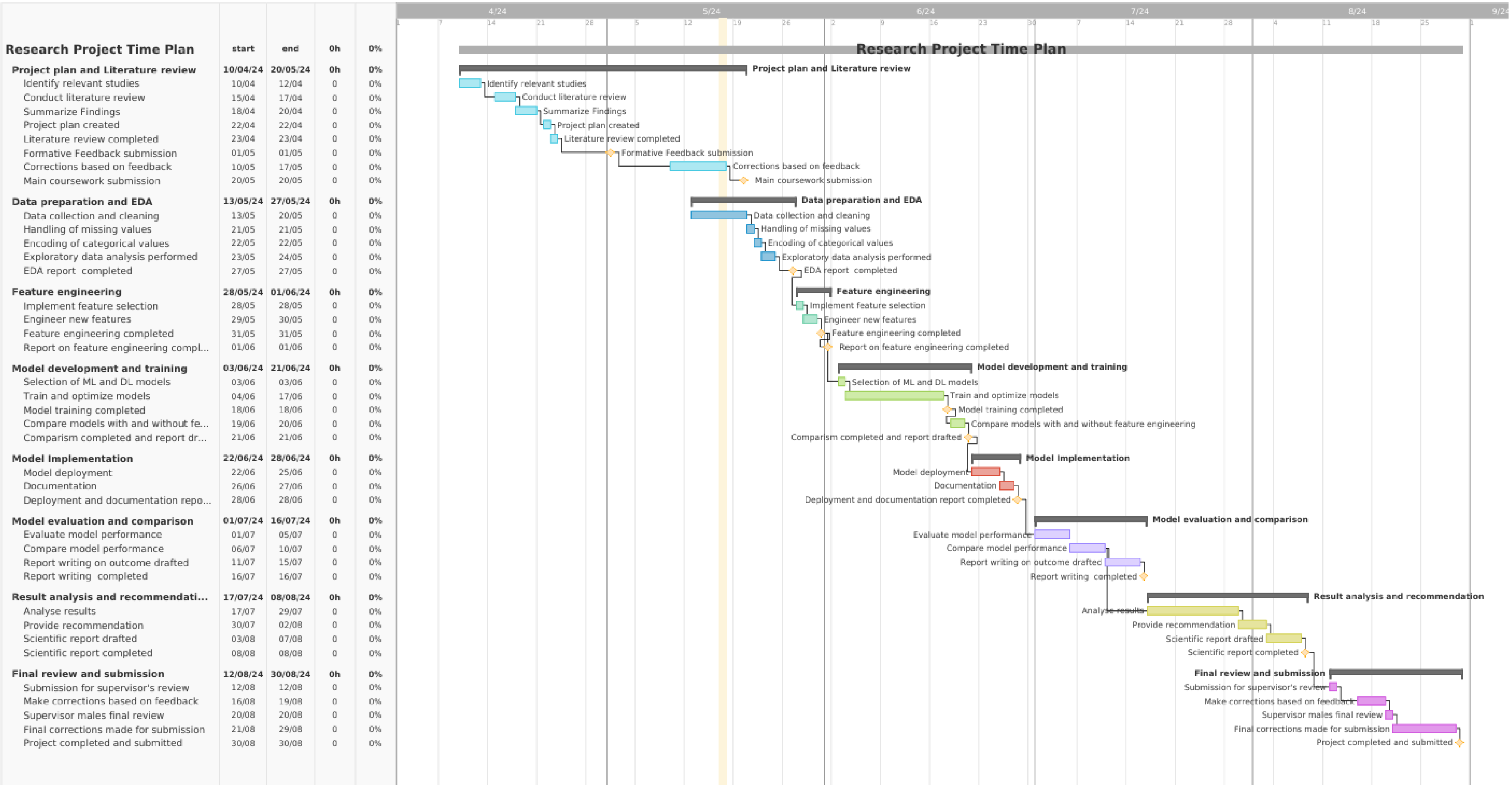
The successful completion of this project could contribute to computer-aided breast cancer diagnosis by capitalising on the strengths of histopathological and mammographic imaging. This research aims to develop a system that could potentially reduce false positives and negatives, leading to earlier and more reliable breast cancer detection. Such advancements could have profound implications for patient outcomes, healthcare resource allocation, and understanding of breast cancer progression (Okorie, et al. 2024).

The project's main contribution is using supervised learning to engineer features for multimodal breast cancer diagnosis. This provides new insights into how powerful features can distinguish different imaging modalities, paving the way for more reliable, understandable, and clinically relevant diagnostic tools (Arif, et al. 2023; Leskovec, et al. 2020).

Time plan

The project is expected to be completed by August 2024. The Gantt chart below provides a visual representation of the timeline and milestones. The project plan is meticulously structured to cover critical tasks over several months, ensuring a comprehensive approach. It begins with a thorough literature review and project planning phase, providing a solid research foundation. After ensuring high-quality datasets and initial insights through data preparation and exploratory data analysis, feature engineering will be conducted to create relevant and impactful features for model development. Substantial time from early June to mid-June is allocated for the project's core, which is model development and training, to ensure a thorough exploration of various algorithms and optimisation techniques. The model implementation, evaluation, and comparison phases are scheduled from June 22 to July 16, ensuring rigorous testing and selecting the best-performing model. Then, the final stages of the project, which include report writing, results analysis, and recommendations, span from July 17 to August 8, culminating in a comprehensive final review and submission on August 30.

The time plan can be adjusted to accommodate unforeseen challenges or developments as the project progresses. For example, potential changes to improve the plan's efficiency and robustness include introducing overlapping tasks to help optimise time. Additionally, incorporating buffer days into critical phases can accommodate unforeseen delays and challenges while keeping the project on track. Regular supervisor check-ins will be held during key phases to identify and address issues early, potentially reducing the need for extensive revisions later in the project.



Research Project Time Plan	start	end	0h	0%
Project plan and Literature review	**10/04/24**	**20/05/24**	**0h**	**0%**
Identify relevant studies	10/04	12/04	0	0%
Conduct literature review	15/04	17/04	0	0%
Summarize Findings	18/04	20/04	0	0%
Project plan created	22/04	22/04	0	0%
Literature review completed	23/04	23/04	0	0%
Formative Feedback submission	01/05	01/05	0	0%
Corrections based on feedback	10/05	17/05	0	0%
Main coursework submission	20/05	20/05	0	0%
Data preparation and EDA	**13/05/24**	**27/05/24**	**0h**	**0%**
Data collection and cleaning	13/05	20/05	0	0%
Handling of missing values	21/05	21/05	0	0%
Encoding of categorical values	22/05	22/05	0	0%
Exploratory data analysis performed	23/05	24/05	0	0%
EDA report completed	27/05	27/05	0	0%
Feature engineering	**28/05/24**	**01/06/24**	**0h**	**0%**
Implement feature selection	28/05	28/05	0	0%
Engineer new features	29/05	30/05	0	0%
Feature engineering completed	31/05	31/05	0	0%
Report on feature engineering compl...	01/06	01/06	0	0%
Model development and training	**03/06/24**	**21/06/24**	**0h**	**0%**
Selection of ML and DL models	03/06	03/06	0	0%
Train and optimize models	04/06	17/06	0	0%
Model training completed	18/06	18/06	0	0%
Compare models with and without fe...	19/06	20/06	0	0%
Comparism completed and report dr...	21/06	21/06	0	0%
Model Implementation	**22/06/24**	**28/06/24**	**0h**	**0%**
Model deployment	22/06	25/06	0	0%
Documentation	26/06	27/06	0	0%
Deployment and documentation repo...	28/06	28/06	0	0%
Model evaluation and comparison	**01/07/24**	**16/07/24**	**0h**	**0%**
Evaluate model performance	01/07	05/07	0	0%
Compare model performance	06/07	10/07	0	0%
Report writing on outcome drafted	11/07	15/07	0	0%
Report writing completed	16/07	16/07	0	0%
Result analysis and recommendati...	**17/07/24**	**08/08/24**	**0h**	**0%**
Analyse results	17/07	29/07	0	0%
Provide recommendation	30/07	02/08	0	0%
Scientific report drafted	03/08	07/08	0	0%
Scientific report completed	08/08	08/08	0	0%
Final review and submission	**12/08/24**	**30/08/24**	**0h**	**0%**
Submission for supervisor's review	12/08	12/08	0	0%
Make corrections based on feedback	16/08	19/08	0	0%
Supervisor males final review	20/08	20/08	0	0%
Final corrections made for submission	21/08	29/08	0	0%
Project completed and submitted	30/08	30/08	0	0%

Figure 1: Gantt Chart

CHAPTER 2

LITERATURE REVIEW

Introduction

Machine learning (ML) and deep learning (DL) modalities have recently attracted attention for their capacity to enhance diagnostic precision and efficiency in breast cancer diagnosis (Yaqoob, et al. 2023). Figure 2 below gives an overview of machine learning (Analytics Vidhya, 2021).

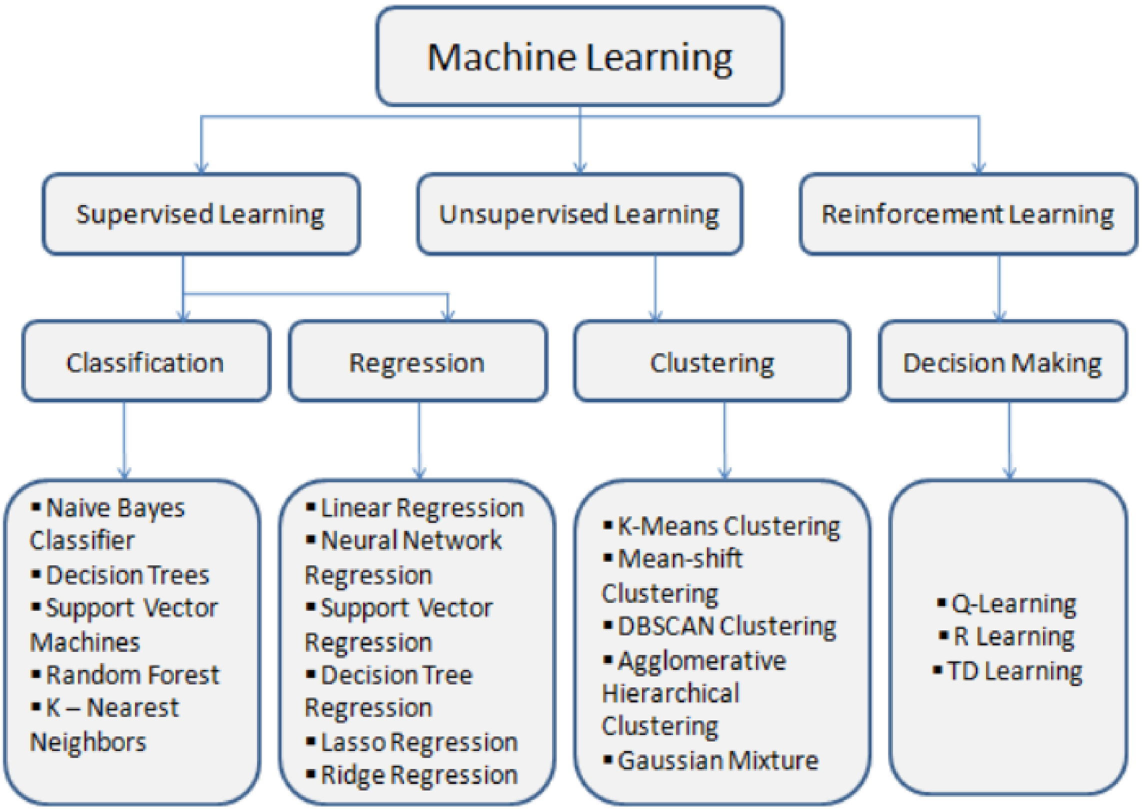


Figure 2: Machine Learning Overview.

Utilising pertinent scholarly sources, this literature survey investigates the applications of machine learning and deep learning. It rigorously assesses the use of deep learning models and machine learning in diagnosing breast cancer, specifically examining their effectiveness, limitations, and possible directions for future advancement. This study will provide valuable insights into the continuous endeavours to enhance early identification, optimise treatment approaches, and elevate the standard of care for persons impacted by this illness.

Sources

The evolution of machine learning (ML) and deep learning (DL) in breast cancer diagnosis reveals a fascinating journey of technological advancement. Yaqoob, et al. (2023) provide an overview of how ML algorithms, including support vector machines (SVM), decision trees, random forests, and convolutional neural networks (CNN), are used to analyse various imaging modalities in breast cancer diagnosis.

Traditional ML algorithms initially laid the groundwork for computer-aided diagnosis. Sakib et al. (2022) examined various ML algorithms using the Breast Cancer Wisconsin dataset, finding that the precision rate achieved was 96.66% and the F1-score was 0.963 by the Random Forest.

Building on this, Ak M. F. (2020) discovered that a simpler Logistic Regression model achieved an impressive 99.12% accuracy, challenging the notion that more complex algorithms always perform better. The field progressed towards more robust methodologies, as evidenced by Awan et al. (2024). Their dual-dataset approach revealed intriguing performance variations across different datasets, with the K-nearest neighbours algorithm achieving a 99% accuracy rate on the Wisconsin breast cancer dataset. Logistic regression performed best on another breast cancer dataset with 83% accuracy. This highlighted the critical influence of dataset characteristics on model performance. Gupta and Garg (2020) further explored the application of SVM, random forests, and artificial neural networks (ANN) on digital mammograms, demonstrating their potential to overcome human interpretation limitations.

As the research community grappled with the complexities of medical imaging data, deep learning emerged as a game-changer. Devi et al. (2024) demonstrated the superiority of DL models, with their Multilayer Perceptron achieving an AUC-ROC of 0.9959 and accuracy of 96.49%, outperforming traditional ML classifiers. This shift towards DL was further solidified by studies like Zhong et al. (2024), who introduced a pioneering multi-task fusion model. Their approach, integrating features from various imaging modalities and clinical data, achieved impressive AUC scores of 0.92 and 0.95 on the CBIS-DDSM and INbreast datasets, respectively.

Sahu et al. (2023) reviewed the potential of DL in breast cancer diagnosis, highlighting studies achieving accuracy as high as 98.94% on the DDSM database. This review underscored the effectiveness of DL-based methods over traditional ML approaches, particularly in feature extraction and classification tasks. Bi et al. (2019) noted that CNN has been employed in medical image analysis since the early 1990s, demonstrating the long-standing potential of these techniques.

Recognising the complex nature of breast cancer manifestation, researchers began exploring multimodal approaches. Del Corso et al. (2024) conducted the groundbreaking P.I.N.K study, creating the first ABVS+DBT dataset for radiomic analysis. Their adaptive ML approach achieved an AUC-ROC of 89.9% for DBT and 85.8% for ABVS using only a few key features, demonstrating the power of multimodal integration.

This trend toward multimodal analysis was further advanced by a collaborative effort from Kayikci and Khoshgoftaar (2023), who proposed a deep learning approach combining mammograms and ultrasound images. Their CNN architecture integrated features from both modalities, yielding enhanced diagnosis accuracy compared to individual modalities. Li et al. (2023) also contributed to this collaborative effort by proposing an ensemble classification approach using a hybrid dimensionality reduction method to extract relevant features from mammogram images, further enhancing diagnostic accuracy.

As the field progressed, researchers began tackling the challenge of limited labelled data in medical imaging. Kim et al. (2023) and Ayana et al. (2021) explored transfer learning techniques, demonstrating how pre-trained models could be fine-tuned for breast cancer diagnosis, achieving high accuracy while reducing computational requirements. Harrison et al. (2023) further Explored the implementation of transfer learning-based deep neural networks to analyse breast cancer histopathology images, achieving top-notch performance in tumour detection and classification.

Yanan Du et al. (2023) researched the differential diagnosis of benign and malignant breast lesions using a deep-learning model based on multimodal images, achieving an AUC of 0.943. Thangavel et al. (2024) introduced a pioneering deep learning-based approach focusing on digital mammogram-based feature extraction and early-stage identification, demonstrating exceptional performance metrics with 99% accuracy.

The latest advancements in the field showcase increasingly sophisticated approaches. Kong et al. (2024) introduced a dual-attention mechanism for feature extraction from ultrasound videos, while Huo et al. (2024) proposed the HiFuse network for capturing features at multiple scales. Do et al. (2024) explored information fusion from different imaging modalities using optimisation techniques and transfer learning with VGG19.

Ray et al. (2024) conducted a comparative study on various pre-trained models for histopathological image analysis, providing valuable insights into the feature extraction capabilities of different architectures. Sahu et al. (2024) combined mammogram and ultrasound images, using transfer learning to overcome limited data. Yaqub et al. (2024) proposed a two-stage approach using Atrous Convolution and multi-scale DenseNet for mammography analysis, showcasing innovative feature extraction techniques. Zakareya et al. (2023) proposed a new deep-learning model using granular computing, learnable activation functions, and an attention mechanism, achieving high accuracy on ultrasound and breast histopathology images.

These studies show rapid advancement in AI-assisted breast cancer diagnosis, from simple ML algorithms to complex, multimodal deep learning approaches. They highlight the field's ongoing challenges and opportunities and point towards a future where AI could revolutionise breast cancer detection and diagnosis.

Critical Evaluation

The literature analysis highlights the considerable potential of machine learning (ML) and deep learning (DL) approaches. Several studies have investigated integrating different imaging modalities, including mammography, ultrasonography, and magnetic resonance imaging (MRI), to use each modality's complementary information; other sources touched on ANN and MLP. In this section, each of these sources will be critically evaluated, focusing on their strengths and limitations, to further understand the research carried out by the above sources.

Traditional Machine Learning Algorithms:

The application of traditional machine learning algorithms to breast cancer detection, as exemplified by Gupta and Garg (2020), Awan et al. (2024), Sakib et al. (2022), and Ak, M.F (2020), reveals a landscape with challenges that extend beyond the mere pursuit of improved accuracy metrics. While contributing to the field, these studies inadvertently highlight several issues that demand attention from the research community.

A fundamental concern lies in the justification for algorithm selection. Gupta and Garg (2020), for instance, offer a comparative analysis of six algorithms but fail to provide a substantive rationale for their choices in the context of breast cancer diagnosis. Their selection is driven more by algorithmic popularity than by a cogent understanding of the unique challenges of medical imaging data. This approach reflects a broader, problematic trend in the field: the prioritisation of algorithmic variety over tailored, problem-specific solutions.

Feature selection, an essential aspect of model development, receives inadequate attention across these studies. The work of Awan et al. (2024) inadvertently highlights this issue through their dual-dataset approach. The significant performance discrepancies observed between the Wisconsin breast cancer dataset (WBCD) and the Breast cancer dataset underscore the profound impact of feature selection on model performance. However, their analysis falls short of providing meaningful insights into the causal factors behind these differences or proposing systematic approaches to address them. This represents a missed opportunity to advance the understanding of feature relevance in breast cancer detection and raises questions about the field's depth of engagement with fundamental machine learning principles.

The presentation of performance metrics in these studies often proves inadequate for clinical contexts. Sakib et al. (2022) emphasise accuracy, reporting a 96.66% figure for random forest. However, this focus obscures more clinically relevant metrics such as sensitivity and specificity. In the high-stakes context of breast cancer diagnosis, where false negatives can have life-threatening consequences, this oversight transcends academic concern and enters the realm of potential clinical danger. The field's strong emphasis on accuracy as the primary success metric reveals a concerning disconnect from actual clinical situations and ethical considerations.

Furthermore, all four studies demonstrate significant limitations in addressing the complexity inherent in medical imaging data. While comprehensive in its algorithmic comparisons, Ak, M.F. (2020)'s work fails to adequately address how these methods cope with the high-dimensionality and intricate spatial relationships in mammographic images. This omission raises questions about the real-world applicability of the results and highlights a critical gap between current research methodologies and the complex requirements of clinical practice. Additionally, overfitting in machine learning is not adequately addressed in these studies. The high accuracies reported by Gupta and Garg (2020), particularly for logistic regression (98.1%), raise significant concerns about potential overfitting. Their limited discussion of cross-validation techniques and the absence of independent test sets weaken confidence in the generalisability of their findings. This issue indicates a more significant trend in the field where there is a preference for impressive accuracy figures over the development of robust, generalisable models capable of performing consistently across different datasets.

Another common flaw across all four studies is the failure to place findings within the larger context of breast cancer diagnosis. The absence of comparative analysis with traditional clinical methods regarding accuracy, speed, and cost-effectiveness leaves a significant gap in understanding the actual value of these computational approaches in real-world clinical settings. This lack of contextualisation is an important oversight in the current research paradigm and limits the potential for meaningful application.

All four studies have inadequately investigated the ethical implications of implementing AI in breast cancer diagnosis. They fail to adequately address the challenges of model explainability, accountability, or potential biases. In the context of life-altering medical decisions, this oversight is not just a theoretical concern but a pressing ethical obligation. The reluctance to address these ethical considerations reflects a disconnect between technological progress and responsible implementation.

Future research must go beyond the current paradigm of incremental improvements in accuracy measures to advance the field meaningfully. Instead, it should focus on developing robust, interpretable models capable of handling the heterogeneity of real-world clinical data. This requires a fundamental shift in research priorities towards developing innovative feature selection methods, creating new performance metrics, integrating domain-specific knowledge, rigorously investigating model interpretability, and establishing thorough ethical frameworks for AI in healthcare settings.

Deep Learning Techniques:

Breast cancer imaging has seen significant advancements with deep learning techniques. Recent work has propelled this progress, as evidenced by studies from Kim et al. (2023), Kong et al. (2024), Thangavel et al. (2024), Huo et al. (2024), Du et al. (2024), Do et al. (2024), and Bi et al . (2019). These studies collectively represent a significant leap forward in handling complex, high-dimensional medical imaging data, each building upon previous work to address unique challenges in the field.

The evolution of feature extraction techniques is an essential theme in this progression. Kong et al. (2024) MV-STCNet introduces a dual-attention mechanism for feature extraction from ultrasound videos, marking a notable advancement in feature engineering. While innovative, its application is limited to ultrasound data, prompting questions about adapting such mechanisms for other modalities or proper multimodal analysis.

Building on the importance of feature extraction, Huo et al. (2024) propose the HiFuse network, addressing the critical need to capture features at multiple scales. This approach shows promise for breast cancer diagnosis, where abnormalities can manifest at various scales. However, the study doesn't explore its application across different imaging modalities, highlighting a recurring challenge: the need for versatile approaches to integrate information from diverse image types.

Addressing this challenge of multimodal integration, Do et al. (2024)’s work on medical image fusion using VGG19 and transfer learning attempts to combine information from multiple imaging modalities. This study represents a step towards more comprehensive diagnostic approaches. However, the reliance on a general-purpose model raises questions about extracting domain-specific features needed for accurate diagnosis.

The theme of model architecture comparison is evident in Ray et al.'s (2024) study of pre-trained models for histopathological image analysis. Their finding of ResNet50's superior performance contributes valuable insights to the ongoing debate about optimal architectures for medical imaging tasks. Yet, the study's limitation to histopathological images underscores the need for more comprehensive evaluations across multiple imaging modalities.

Researchers face the challenge of limited annotated datasets as the field progresses. Kim et al. (2023) address this by implementing a CNN-based ideal model observer via transfer learning for multi-slice simulated breast CT images. While innovative, this approach raises important questions about the validity of transferring features from non-medical fields to specialised medical imaging, highlighting the tension between leveraging existing knowledge and ensuring domain-specific accuracy.

Thangavel et al. (2024) further illustrates the field's move towards integrated approaches, combining pre-trained ResNet and U-Net models for feature extraction and segmentation of mammographic images. Their focus on early-stage identification addresses a critical clinical need. However, using pre-trained models not optimised for mammographic images' unique characteristics reveals an ongoing challenge: balancing computational efficiency with domain-specific optimisation.

Du et al. (2024) tackles another frontier in the field: multimodal data integration. By combining mammography and MRI sequences, their work represents a step towards more comprehensive diagnostic tools. However, the lack of analysis of computational requirements and comparison with traditional methods highlights a recurring theme: the need to balance advanced techniques with practical clinical considerations.

As the field advances, it's becoming increasingly clear that ethical considerations and rigorous validation must accompany technical progress. A standard limitation across these studies is the inadequate attention to the ethical implications of AI in breast cancer diagnosis, including issues of algorithmic bias and data privacy. The work of Bi et al. (2019) stands out in this regard, providing valuable insights into these often neglected aspects and highlighting the need for a more holistic approach to AI in oncology.

The narrative of deep learning in breast cancer imaging is one of rapid progress and persistent challenges. While these studies demonstrate significant potential, they also highlight the need for more comprehensive evaluations that address accuracy, interpretability, ethical implications, and clinical relevance. As the field moves forward, researchers must prioritise technical advancements, reproducibility, and real-world applicability, ensuring that the promise of AI translates into tangible improvements in breast cancer diagnosis and patient care.

Multimodal Approaches and Data Fusion

As breast cancer imaging evolves, researchers increasingly explore multimodal approaches to enhance diagnostic accuracy. This shift represents a natural progression from single-modality deep learning techniques to more comprehensive diagnostic frameworks.

Sahu et al. (2024)'s work combining mammograms and ultrasound images exemplifies this trend. Their pragmatic use of transfer learning to overcome limited data is noteworthy, yet the study needs to explore advanced feature engineering techniques specific to these modalities more in-depth. The potential synergy between mammographic and ultrasound features remains to be explored, highlighting a recurring challenge in multimodal research.

Yaqub et al. (2024) showcase innovative feature extraction techniques with their two-stage approach using Atrous Convolution and multi-scale DenseNet for mammography analysis. Atrous Convolution captures multi-scale features without losing spatial resolution, which is needed for detecting subtle abnormalities. However, the study's focus on a single modality limits its applicability in a comprehensive diagnostic framework.

Zakareya et al. (2023) contribute significantly to feature engineering with their granular computing approach. This technique allows for detailed feature extraction by dividing images into smaller regions, capturing localised information critical for distinguishing between benign and malignant cases. Introducing learnable activation functions and attention mechanisms further enhances the model's capability to learn complex features. However, the fixed granule size and increased parameter count introduce new challenges, highlighting the delicate balance between model complexity and performance.

Zhong et al. (2024) propose an innovative fusion of digital breast tomosynthesis (DBT) and automated breast volume scanner (ABVS) data using a density-based gating mechanism. While this approach demonstrates originality in dynamically assigning weight to modality contributions, it lacks robust theoretical justification and comprehensive evaluation. The study's examination of the interplay between DBT and ABVS provides valuable insights but falls short of thoroughly investigating how these modalities capture distinct aspects of tumour biology.

Del Corso et al. (2024) offer a more sophisticated approach to feature selection and model evaluation in their ABVS and DBT data analysis. Their adaptive feature selection method and nested leave-one-out cross-validation strategy address the high dimensionality issue common in medical imaging studies. However, the study does not thoroughly examine how selected features relate to underlying biological processes, limiting clinical interpretability.

Kayikci and Khoshgoftaar (2023) take a different approach by integrating clinical, copy number alteration, and gene expression data. This study represents a promising direction for personalised cancer diagnostics by bridging imaging and molecular data. However, the simplistic feature concatenation approach fails to account for the different scales and distributions of features from disparate data types, potentially undermining the synergies between modalities.

A common limitation across these studies is the inadequate addressing of potential dataset biases. The lack of comprehensive analysis of demographic composition and model performance across diverse populations raises concerns about generalizability and ethical implications, echoing issues observed in earlier deep-learning studies.

These recent works in multimodal breast cancer diagnosis demonstrate significant progress but also highlight persistent challenges. The field is moving towards more comprehensive diagnostic frameworks, yet issues of feature integration, bias mitigation, and clinical interpretability remain. Future research must prioritise stronger theoretical foundations for modality selection, enhanced measures to reduce bias, and increased focus on practical clinical implementation. As the field advances, integrating diverse data types and developing robust, interpretable models will be crucial in realising the full potential of AI in breast cancer diagnosis.

Transfer Learning and Pre-trained Models

As the field progresses towards more sophisticated multimodal approaches, the role of transfer learning in breast cancer imaging has come under scrutiny. Harrison et al. (2023) and Ayana et al. (2021) analyse transfer learning methods, revealing progress and significant gaps in this critical area.

These reviews highlight an issue: the prevalent use of natural image datasets like ImageNet as source domains for medical imaging applications. While showing promise, this approach raises critical questions about domain relevance. Ayana et al. (2021) points out the unique characteristics of medical images, such as grayscale format and intricate textures, which differ substantially from natural images. This domain discrepancy potentially compromises the efficacy of knowledge transfer, echoing earlier concerns about the applicability of general-purpose models in specialised medical contexts.

The authors propose cross-modal transfer learning to address this, using medical imaging modalities like mammography as source domains. This approach has demonstrated superiority over cross-domain transfer from natural images in specific scenarios. However, a thorough analysis of different source domains across various breast cancer imaging tasks is absent, representing a research gap.

Both reviews extensively discuss fine-tuning strategies, with Harrison et al. (2023) providing a more detailed analysis. They note that fine-tuning generally outperforms feature extraction in breast cancer imaging tasks. Yet, the reviews fall short in critically evaluating fine-tuning techniques, such as the optimal number of layers to fine-tune or the impact of learning rate schedules. The challenge of fine-tuning with limited target domain data, a common scenario in medical imaging, is inadequately addressed, mirroring earlier discussions about the scarcity of large, annotated datasets in this field. The potential for negative transfer, where source domain knowledge hinders target task performance, is a critical issue primarily overlooked in both reviews. This oversight is particularly concerning given the significant domain shift between natural and medical images, underscoring the need for robust methods to address negative transfer in breast cancer diagnosis tasks.

Ethical implications of AI in breast cancer diagnosis receive cursory treatment in these reviews, reflecting a persistent gap in the field. While Harrison et al. (2023) mentions the potential for bias in training data, they do not delve into the risks associated with transfer learning from pre-trained models. The use of ImageNet, known to contain various biases (Prabhu and Birhane, 2020), as a source domain could introduce unintended biases in breast cancer diagnosis models. Furthermore, the critical issues of model explainability and accountability, crucial for clinical adoption, are inadequately addressed, echoing earlier concerns about the black-box nature of deep learning models in medical contexts.

Performance Comparison and Review Studies

While comprehensive in their coverage of various AI techniques, the methodological approaches across these studies reveal drawbacks hindering real-world applicability. The heavy reliance on publicly available datasets like DDSM, MIAS, and INbreast, as seen in Sahu et al. (2024), facilitates reproducibility but fails to capture the complex variability inherent in clinical settings. This overreliance potentially leads to overfitting and a false sense of progress, as models may perform exceptionally well on curated datasets but falter when confronted with real-world mammogram heterogeneity.

Li et al. (2023) highlighted the inconsistency in preprocessing steps, data augmentation techniques, and hyperparameter tuning approaches across studies, which undermines the validity of direct comparisons. This lack of standardisation impedes meaningful cross-study analysis and raises questions about the robustness and generalizability of reported results. Establishing standardised data preparation and model evaluation protocols would enable more reliable comparisons and facilitate meta-analyses.

A critical shortcoming is the superficial treatment of performance metrics. While standard metrics like accuracy, sensitivity, specificity, and AUC are routinely reported, there is a conspicuous lack of in-depth analysis of their clinical relevance. Devi et al. (2024) addressed this issue but failed to propose concrete solutions. The binary classification focus (normal vs. abnormal) prevalent in most studies oversimplifies the clinical decision-making process. Moreover, Li et al. (2023) noted that the limited focus on false positive rates overlooks the significant psychological and economic impacts of unnecessary follow-ups.

Inconsistencies in results across studies, particularly regarding the effectiveness of different deep learning architectures and preprocessing steps, reveal a deeper issue. While Sahu et al. (2023) advocate for ResNet-based models, Li et al. (2023) champion DenseNet variants. This discrepancy underscores the need for more rigorous ablation studies and sensitivity analyses to isolate the impact of specific architectural choices and preprocessing steps. A tendency to report incremental improvements hinders the field's progress without thoroughly investigating the underlying factors contributing to performance gains.

A significant oversight is the failure to investigate model interpretability and explainability thoroughly. Although Devi et al. (2024) acknowledges this gap, they fail to examine its implications for clinical adoption and regulatory approval. The black-box nature of many deep learning models poses significant challenges for integration into clinical workflows, yet the reviewed studies offer few innovative solutions to this pressing issue.

Ethical implications of AI in breast cancer diagnosis continue to receive cursory treatment. Li et al. (2023) briefly acknowledges the possibility of AI systems worsening healthcare inequalities but fail to thoroughly examine the ethical consequences of implementing these technologies. The current body of research has primarily overlooked essential issues related to data privacy, informed consent, and the potential for algorithmic bias, echoing concerns raised in earlier studies about the ethical dimensions of AI in healthcare.

Research Gaps and Questions

The literature review offers an overview of the growing field of machine learning (ML) and deep learning (DL) applications in breast cancer diagnosis. It highlights the significant strides made in using diverse imaging modalities, including mammograms, ultrasound, and magnetic resonance imaging (MRI), to enhance diagnostic accuracy and efficiency. Researchers have explored a wide range of approaches, from ensemble learning methods (Li et al., 2023) and transfer learning techniques (Ayana et al., 2021; Harrison et al., 2023) to the application of traditional machine learning algorithms (Awan et al., 2024; Sakib et al., 2022; Ak et al., 2020). As shown by Kayikci and Khoshgoftaar (2023), integrating multiple imaging modalities has shown promise in improving diagnostic performance.

However, this extensive body of research also reveals a critical gap in the field: a comprehensive analysis of advanced feature engineering methods in the context of supervised ML and DL algorithms for breast cancer classification. While the importance of feature extraction and selection has been acknowledged in some studies, there is a noticeable tendency in the literature to focus predominantly on algorithm selection and performance metrics. This focus often comes at the expense of a deeper exploration of feature engineering's crucial role in enhancing model generalisability, diagnostic accuracy, and clinical applicability.

Moreover, the review exposes a significant deficiency in studies that successfully combine sophisticated feature extraction techniques with thorough multimodal analysis. The prevailing approach of treating each imaging modality in isolation before fusion overlooks the potential for extracting cross-modal features that could unveil valuable correlations, potentially missing crucial diagnostic indicators.

Despite some attempts at multimodal integration, the review also highlights a critical shortcoming in the interpretability of extracted features. For ML and DL models to gain widespread clinical adoption, it is paramount that clinicians understand and trust the engineered features. The lack of emphasis on feature interpretability creates a 'black box' scenario, particularly problematic in the high-stakes context of cancer diagnosis. Additionally, longitudinal studies investigating how temporal feature changes across multiple imaging modalities could enhance diagnosis, particularly in monitoring disease progression or treatment response, are rare. This temporal dimension represents a largely untapped resource in the current research landscape.

The identified gaps in current research underscore the pressing need for a study that systematically evaluates advanced feature engineering methods explicitly tailored to breast cancer imaging. This research should seamlessly integrate these methods with sophisticated multimodal fusion techniques. Exploring the synergy between advanced feature engineering techniques and various supervised learning ML and DL algorithms aims to gain valuable insights into the discriminative power of different features and their relevance to breast cancer classification. Central to this research is a thorough evaluation of how feature engineering impacts model performance metrics such as accuracy, precision, recall, and F1-score. Equally crucial is assessing the models' interpretability and explainability, ensuring that the developed tools can provide clinicians with reliable and trustworthy diagnostic support. This comprehensive evaluation is vital for facilitating informed decision-making and improving patient outcomes.

This study's unique contribution lies in its comprehensive approach to feature engineering in multimodal breast cancer diagnosis, which uses only supervised learning approaches. This research aims to provide unprecedented insights into the discriminative power of features across different imaging modalities by systematically evaluating various feature engineering techniques across histopathological and mammographic images. This approach to multimodal feature engineering and fusion represents a significant advance over existing studies focusing on single modalities or employing simplistic feature combination methods.

Based on these identified gaps and the aims and objectives of the project, the following research questions have been formulated:

1. How can advanced feature engineering techniques enhance the classification performance of ML and DL models in breast cancer diagnosis using multimodal imaging data (histopathological and mammographic images)?
2. What are the optimal data fusion methods for integrating multimodal imaging data to improve diagnostic accuracy and efficiency?
3. How can interpretability techniques be integrated into developing machine learning models for breast cancer diagnosis to enhance clinical trust and adoption?

These research questions aim to address the critical gaps identified in the existing literature and align with the project's overarching goal of developing a robust, interpretable, and clinically applicable AI-based diagnostic tool for breast cancer. The findings from this research are expected to contribute significantly to the field of computer-aided diagnosis, potentially improving early detection and patient outcomes while addressing ethical and practical implementation challenges.

Table 1 below summarises the research examined in this literature review. The research is presented according to the model type, the specific tasks carried out, the imaging modalities used, the datasets employed, and the primary approaches adopted by the authors. This table briefly summarises the notable advancements made by various researchers in machine learning and deep learning applications for diagnosing breast cancer.

Table 1: Summary of Reviewed Literature

Author(s)	**Type**	**Task(s)**	**Image**	**Dataset**	**Method**
Yaqoob et al.	Review	Diagnostic accuracy	Mammography, Ultrasound, MRI	Not specified	ML/DL diagnostic overview
Sakib et al.	Single-task Model	Classification	Mammography	Breast Cancer Wisconsin dataset	Random Forest: High precision
Ak M. F.	Single-task Model	Classification	Mammography	Breast Cancer Wisconsin dataset	Logistic Regression: High accuracy
Awan et al.	Single-task Model	Classification	Mammography, Ultrasound	Breast Cancer, Wisconsin datasets	KNN: Top accuracy
Gupta and Garg	Single-task Model	Classification	Mammography	Digital Mammograms	Ensemble methods outperform
Devi et al.	Single-task Model	Classification	Mammography	Breast Cancer Wisconsin dataset	MLP: Best performance
Zhong et al.	Multi-task Fusion Model	Classification, Segmentation	Mammograms, Ultrasounds	CBIS-DDSM, INbreast	Fusion with MLP/Residual net
Sahu et al. (2023)	Multi-task Model	Segmentation, Classification	Mammograms	DDSM, MIAS	Used ResNet, U-Net
Sahu et al. (2024)	Multi-task Model	Classification	Mammograms, Ultrasound	Not specified	Transfer learning

Del Corso et al.	Single-task Model	Classification	Mammography, ABVS, DBT	ABVS+DBT dataset	Adaptive feature selection
Bi et al.	Single-task Model	Classification	Mammography	Not specified	CNNs evolution in diagnosis
Kayikci and Khoshgoftaar	Fusion Model	Classification	Mammography, Ultrasound	Not specified	CNN: Combined features
Li et al.	Ensemble Model	Classification	Mammography	Digital Mammograms	Ensemble: Multiple classifiers
Kim et al.	Single-task Model	Classification	CT	Simulated Breast CT data	CNN: Transfer learning
Ayana et al. (2021)	Transfer Learning Model	Classification	Mammography, Ultrasound, Clinical data	Not specified	Fine-tuned models (VGG, ResNet)
Ayana et al. (2024)	Transfer Learning Model	Classification	Not specified	Not specified	Transfer learning specifics
Harrison et al.	Transfer Learning Model	Classification	Histopathology	Not specified	Transfer learning: Deep nets
Thangavel et al.	Single-task Model	Classification, Segmentation	Mammography	Not specified	ResNet/U-Net integration
Kong et al.	Single-task Model	Feature Extraction	Ultrasound	Not specified	Dual-attention features
Huo et al.	Multi-task Model	Multiscale Feature Fusion	Mammography, Histopathology	Not specified	HiFuse for fusion

Do et al.	Multi-task Fusion Model	Image Fusion, Classification	Mammography, Ultrasound	Not specified	Fusion with VGG19
Ray et al.	Comparative Study	Feature Extraction	Histopathology	Not specified	ResNet50 for extraction
Yaqub et al.	Multi-task Model	Feature Extraction, Classification	Mammography	Not specified	Atrous/DenseNet
Zakareya et al.	Single-task Model	Classification	Ultrasound, Histopathology	Not specified	Granular computing with attention

CHAPTER 3

NEW IDEAS & APPROACH

Introduction

This chapter presents an original approach and a methodology for breast cancer diagnosis using multimodal imaging data. The framework builds on the critical evaluation and addresses the limitations and research gaps identified in the literature review (Chapter 2). By combining advanced supervised machine learning techniques with a sophisticated multimodal feature engineering pipeline, the approach aims to capitalise on the complementary strengths of histopathological and mammographic imaging data to enhance diagnostic precision and reliability.

The methodology for this study is distinguished by its synergistic treatment of feature engineering and data fusion, aspects that have been largely overlooked in prior research. The goal is to systematically evaluate various feature engineering techniques across both imaging modalities and employ an unconventional data fusion strategy. This strategy promises to provide unprecedented insights into the discriminative power of features across different imaging modalities, allowing for more clinically applicable AI-based diagnostic tools.

Methodology

Figure 3 below gives an overview of an encompassing methodology employed. It will span data acquisition, preprocessing, feature engineering, model development, and evaluation. Five supervised learning models will be compared to address the study's aims and objectives and to bridge the gaps identified in the literature review.

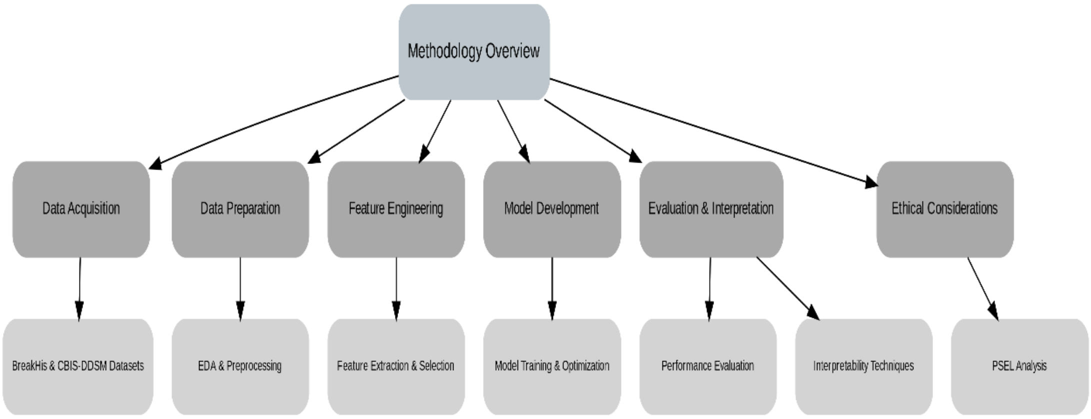


Figure 3: Methodology Flowchart

Data Acquisition

The study utilises two complementary datasets: BreakHis for histopathological images and CBIS-DDSM for mammographic images. The BreakHis dataset comprises 7,909 microscopic images of breast tumour tissue from 82 patients at various magnification factors (40X, 100X, 200X, and 400X). These images, categorised into benign and malignant classes, provide detailed textural and morphological information at the cellular level (Mehta, et al. 2023). The CBIS-DDSM dataset contains 10,237 mammographic images with annotations for benign and malignant lesions, offering a panoramic representation of anatomical characteristics important for accurate diagnosis (Feng, et al. 2023). Using these datasets allows for multi-scale analysis of breast cancer, addressing the limitations of single-modality approaches identified in previous studies (Nguyen et al. 2024). A preprocessing pipeline is implemented, including image standardisation, data augmentation, and robust error handling. Detailed implementation of these procedures is presented in Chapter 4.

Hardware Specifications and Computational Resources

This research uses the enhanced computational resources provided by Google Colab Pro+, specifically the A100 GPUs, each offering 80GB of RAM. These high-performance GPUs enable us to train and evaluate complex multimodal models efficiently. The exact GPU model may vary based on availability, but Colab Pro+ consistently provides access to these advanced computational resources. This environment allows for the implementation of sophisticated simultaneous feature extraction techniques across histopathological and mammographic images, handling the computational demands of our approach while maintaining flexibility and scalability.

Exploratory Data Analysis (EDA)

The exploratory data analysis will be a critical first step in understanding the characteristics of both the BreakHis and CBIS-DDSM datasets. This process will inform the subsequent preprocessing and modelling decisions, ensuring the data is well understood before any transformations are applied.

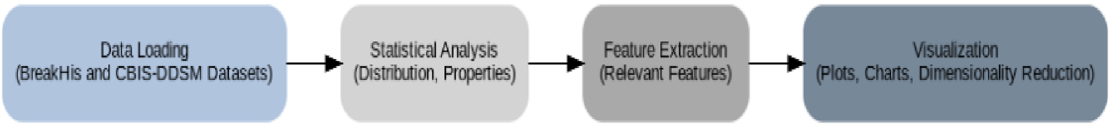

Data Loading
(BreakHis and CBIS-DDSM Datasets)
Statistical Analysis
(Distribution, Properties)
Feature Extraction
(Relevant Features)
Visualization
(Plots, Charts, Dimensionality Reduction)

Figure 4: EDA Workflow

As shown in Figure 4, for the BreakHis dataset, the EDA will examine the distribution of benign and malignant samples across different magnification levels (40X, 100X, 200X, and 400X). This analysis will be crucial in identifying any class imbalances the modelling approach must address. Image quality across these magnification levels will also be investigated, looking for potential issues such as blur, noise, or artefacts that could impact the model's performance.

Key features will be extracted and analysed to gain deeper insights into the image characteristics. These will include texture features using the Grey-Level Co-

Occurrence Matrix (GLCM), shape features using Local Binary Patterns (LBP), and colour features using colour histograms. Examining the distributions of these features aims to identify potential discriminative patterns between benign and malignant samples. Additionally, an exploration of the relationship between these features and the magnification levels will be conducted to understand how the level of detail impacts the distinguishing characteristics of the samples.

For the CBIS-DDSM dataset, the analysis will focus on the distribution of different lesion types (masses and calcifications) within the benign and malignant classes. This will provide insights into the diversity of the dataset and potential challenges in distinguishing between different lesion types. The distribution of breast density categories will also be examined, as this factor can significantly impact lesion visibility and, consequently, model performance.

Metrics such as contrast and noise levels will be calculated and analysed to assess image quality in the CBIS-DDSM dataset. This analysis will be crucial in informing preprocessing steps and potentially guiding model design decisions to account for variability in image quality. The relationship between these quality metrics and the pathology classifications will also be examined to identify potential correlations. For both datasets, dimensionality reduction techniques such as t-SNE (t-Distributed Stochastic Neighbour Embedding) and UMAP (Uniform Manifold Approximation and Projection) will be employed to explore the high-dimensional feature space. These techniques will help identify potential clusters or patterns in the data that may not be apparent through traditional statistical analyses.

Statistical analyses will play a key role in the EDA process. Chi-square tests for categorical variables and ANOVA tests for continuous variables will be conducted to assess their associations with the target variable (benign/malignant). Effect size calculations, such as Cohen's d, will quantify the magnitude of differences between benign and malignant samples across various features. These analyses will provide quantitative support for observations and guide the feature selection process. Finally, correlation analyses between different features and metadata elements will be performed to identify potential multicollinearity issues and inform the feature selection strategy for model development.

Data Preprocessing

Based on the insights gained from the exploratory data analysis, A thorough preprocessing pipeline will be implemented to prepare the data for advanced machine learning models. Figure 5 below shows an overview of the data preprocessing steps.

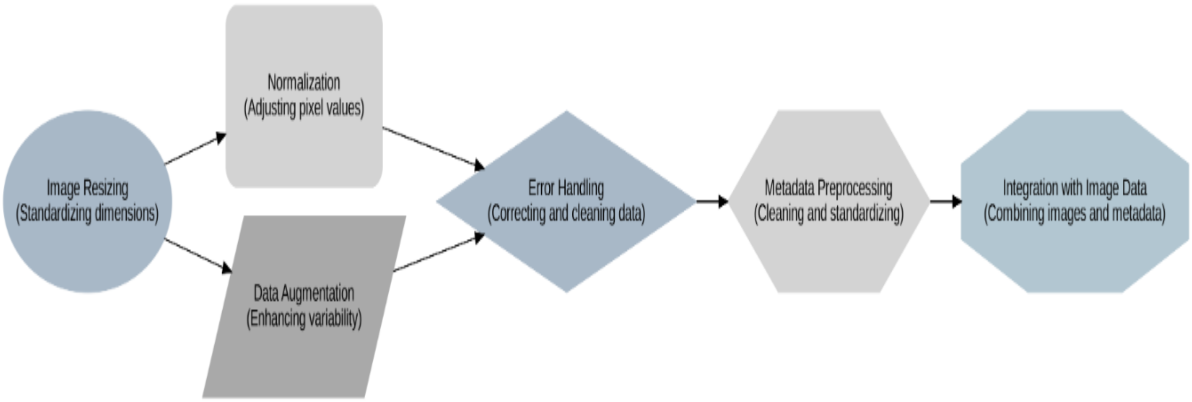


Image Resizing
(Standardizing dimensions)
Normalization
(Adjusting pixel values)
Data Augmentation
(Enhancing variability)
Error Handling
(Correcting and cleaning data)
Metadata Preprocessing
(Cleaning and standardizing)
Integration with Image Data
(Combining images and metadata)

Figure 5: Data Preprocessing Workflow

Image standardisation will be a crucial first step. All images from both datasets will be resized to a uniform resolution of 224 x 224 pixels, balancing the need for detailed features with computational efficiency. This resolution has been chosen to align with the input requirements of the pre-trained convolutional neural network architectures that will be utilised. Following resizing, pixel values will be normalised to the range [0, 1] using Min-Max scaling, ensuring consistent input ranges across all images.

A duplicate removal strategy will be implemented using image hashing algorithms to address potential data redundancy and biases. This step is vital for preventing potential biases in model training and evaluation arising from overrepresenting certain images. Each image will be assigned a unique hash based on its pixel content, allowing us to identify and remove duplicates efficiently.

Data augmentation techniques will be employed to enhance model robustness and mitigate overfitting. Our augmentation strategy will include random rotations, width and height shifts, zoom adjustments, horizontal flips, and brightness variations. The EDA findings and domain knowledge of typical variations in medical imaging will inform the specific ranges for these augmentations (Chen, et al. 2024).

A careful cleaning and imputation strategy will address the handling of missing data, particularly in the CBIS-DDSM metadata. Multiple Imputation by Chained Equations (MICE) for continuous variables will be used to preserve the dataset's statistical properties. Mode imputation will be used where appropriate for categorical variables like breast density, and records with critical missing data that cannot be reliably attributed will be flagged for potential exclusion from the analysis (Zhang et al. 2023). Robust error handling mechanisms are used throughout the preprocessing pipeline, including logging processing errors, such as corrupt image files or inconsistent metadata. These logs will be regularly reviewed to identify any systematic issues in the data that may require further investigation or adjustment of our preprocessing steps.

Data Partitioning

To ensure a rigorous and unbiased evaluation of our models, a stratified sampling approach to partition both datasets into training (70%), validation (15%), and test (15%) sets will be employed. This stratification is crucial, as it maintains the original distribution of benign and malignant cases across all subsets, mitigating the risk of sampling bias that could skew our results and addressing the limitations of previous studies that lacked proper data partitioning (Wang, et al. 2023).

For the BreakHis dataset, an additional constraint in the partitioning strategy is to ensure that images from the same patient do not appear in different subsets. This patient-wise splitting is essential to prevent data leakage and provide a more realistic assessment of our models' performance on unseen data.

The CBIS-DDSM dataset is stratified based on the pathology (benign/malignant) and the type of lesion (mass/calcification), ensuring a balanced representation of these critical factors across all subsets. This approach will enable our models to learn and be evaluated on a diverse range of mammographic presentations of breast cancer. Class weights will be calculated and applied during training to address the potential imbalance in class distribution. These weights will be inversely proportional to the class frequencies, ensuring that our models pay equal attention to benign and malignant cases, regardless of their relative prevalence in the dataset.

Multimodal Feature Engineering Framework

An advanced, multi-stage feature engineering pipeline addresses the research gap regarding the lack of comprehensive feature engineering analysis in breast cancer diagnosis. The feature engineering pipeline extracts extensive discriminative features from histopathological and mammographic images. This process aims to capture subtle tissue characteristics and image patterns that may not be immediately apparent in the raw pixel data, potentially enhancing the diagnostic capabilities of our models. The feature engineering framework consists of four main stages, as shown in Figure 6:

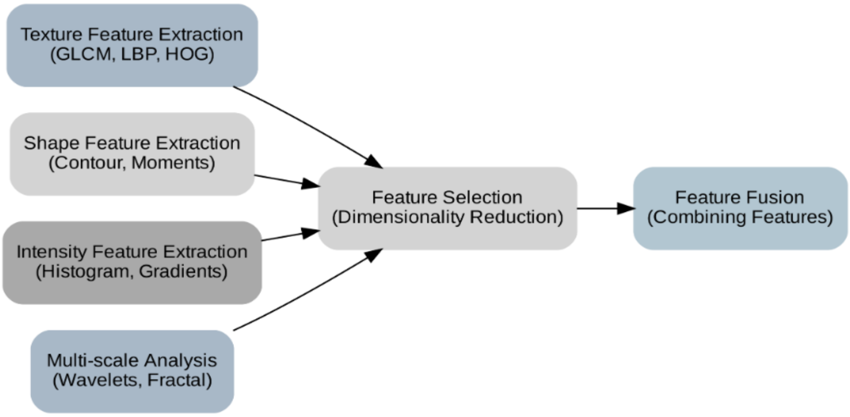

Texture Feature Extraction (GLCM, LBP, HOG)
Shape Feature Extraction (Contour, Moments)
Intensity Feature Extraction (Histogram, Gradients)
Multi-scale Analysis (Wavelets, Fractal)
Feature Selection (Dimensionality Reduction)
Feature Fusion (Combining Features)

Figure 6: Feature Engineering

A multifaceted feature extraction approach will be implemented for the histopathological images from the BreakHis dataset. This approach will begin by computing Gray Level Co-occurrence Matrix (GLCM) features, which provide valuable information about the texture patterns in the tissue samples. These features include contrast, correlation, energy, and homogeneity, offering insights into the spatial relationships between pixels at different grey levels.

Next, local binary patterns (LBP) will be extracted. LBP is a powerful texture descriptor that captures local spatial patterns in images. It is beneficial for identifying micro-textures in histopathological images, which can indicate cellular abnormalities associated with malignancy. A uniform LBP with a radius of 1 and 8 sampling points has performed well in previous histopathological studies.

Haralick texture features will also be computed, providing 14 textural features derived from the GLCM. These features, including angular second moment, contrast, correlation, and entropy, offer a comprehensive characterisation of the image texture that has proven valuable in distinguishing between benign and malignant tissue structures.

Lastly, Histogram of Oriented Gradients (HOG) features will be extracted from the histopathological images. HOG captures the distribution of gradient orientations in localised portions of the image, providing information about cellular structures' shape and edge characteristics. This can be particularly useful in identifying the irregular cell shapes often associated with malignancy (Huang, et al. 2023).

The feature engineering pipeline will be tailored to capture the unique characteristics of the mammographic images from the CBIS-DDSM dataset. First, HOG features will be extracted, which are particularly effective in capturing mammograms' shape and edge information of masses and calcifications (Li et al., 2024).

A Wavelet Transform will then be applied to the images, decomposing them into multiple levels of approximation and detail coefficients. This multi-resolution analysis can reveal subtle textures and patterns at different scales, which is crucial for detecting the variety of abnormalities that can appear in mammographic images.

To capture the complexity of tissue patterns, the Fractal Dimension of the images will be computed using the box-counting method. Fractal analysis has shown promise in characterising the irregular structures often associated with malignant lesions in mammograms (Ahmed, et al. 2023).

Texture Entropy features will also be extracted to measure the randomness or unpredictability in the image textures. This can be particularly useful in distinguishing between the more organised structures of normal tissue and the chaotic patterns often seen in malignant lesions. This step aims to provide a more comprehensive representation of the underlying data characteristics, potentially improving model performance (Patel and Kashyap 2023).

After extracting these features, a robust standardisation process will be applied using StandardScaler. This step ensures that all features contribute equally to the subsequent machine learning models regardless of their original scales. The risk of certain features dominating the learning process simply due to their larger magnitude is mitigated by centering the features around zero and scaling them to unit variance. The goal is to enhance the overall discriminative power of the feature set and address the research gap in effective feature fusion methods for diverse imaging modalities (Xie, et al. 2024).

Model Development and Architecture

As illustrated in Figure 7, the study will encompass various machine learning models, from traditional algorithms to advanced deep learning architectures. This multi-model approach will allow for a comprehensive evaluation of the impact of the feature engineering efforts and the potential benefits of multimodal learning.

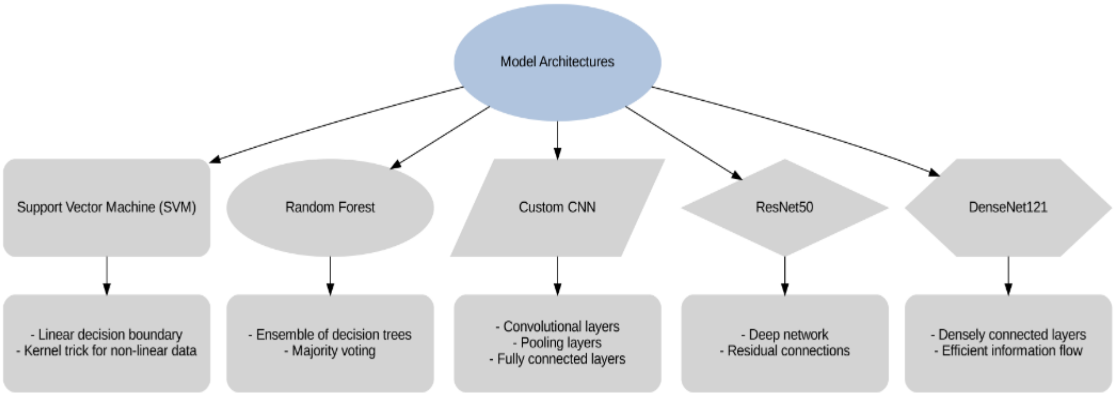


Figure 7: Model Architecture Comparison

The proposed methodology will begin with developing single-modality models using only the BreakHis dataset. This approach will establish a strong foundation for comparing and understanding different methodologies' effectiveness on the specific dataset. It encompasses a range of models, from traditional machine learning algorithms to advanced deep learning architectures. Support Vector Machines (SVM), Random Forests, custom Convolutional Neural Networks (CNN), and transfer learning approaches using ResNet50 and DenseNet121. A multimodal fusion approach is introduced to use complementary information from histopathological and mammographic images.

Traditional machine learning models focus on effective feature engineering and selection; these models serve as baselines and provide insights into the importance of hand-crafted features. The deep learning approaches include a custom CNN architecture explicitly designed for image classification and transfer learning implementations using pre-trained ResNet50 and DenseNet121 models.

The multimodal fusion approach introduces a dual-stream architecture that simultaneously processes histopathological and mammographic images. This innovative design aims to capture complementary information from both imaging modalities, potentially enhancing diagnostic accuracy. Chapter 4 elaborates on each model's specific architectural details and implementation strategies.

Building upon these single modality models, multimodal architectures incorporating histopathological and mammographic data will be developed. These dual-stream models will include:

Custom Dual-Stream CNN: This model will consist of two parallel CNN streams, one for histopathological images and one for mammographic images. Each stream will have a similar architecture to the single modality CNN, with the outputs concatenated before the final classification layers.

Dual-Stream ResNet50: The ResNet50 architecture will be modified to accept inputs from both imaging modalities. The features extracted by the ResNet50 base for each modality will be concatenated before the final fully connected layers.

Dual-Stream DenseNet121: Similar to the dual-stream ResNet50, the DenseNet121 architecture will be adapted to simultaneously process both types of images. Finally, a novel multimodal model will be developed that incorporates both the raw image data and the engineered features from both modalities. This custom dual-stream CNN with feature fusion will have separate inputs for the raw images and the engineered features, with a custom architecture designed to combine these diverse inputs optimally.

Model Training and Optimisation

All models will be trained using the Adam optimiser, which is chosen for its adaptive learning rate capabilities and often leads to faster convergence. Binary cross-entropy will be used for the loss function, which is appropriate for our binary classification task (benign vs. malignant). Figure 9 shows the training and evaluation process.

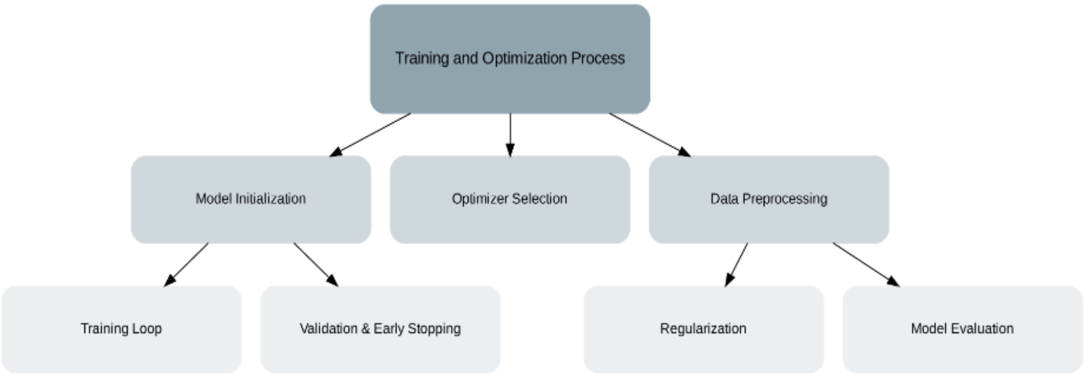


Figure 8: Training and Optimisation Flowchart

An early stopping mechanism with a patience of 5 epochs will be implemented to prevent overfitting and monitor the validation loss. This approach will allow for halting training when the model's performance on the validation set stops improving, helping to ensure good generalisation to unseen data.

Class weights will be computed and applied during the training process to address the class imbalance in the datasets. These weights will be inversely proportional to the class frequencies, ensuring that the models pay equal attention to benign and malignant cases, regardless of their relative prevalence in the dataset. L2 regularisation with a factor of 0.001 will be employed for the deep learning models to mitigate overfitting further. Additionally, dropout layers with a rate of 0.5 will be used in the custom CNN architectures, introducing randomness during training that helps prevent the model from relying too heavily on any features.

Finally, all models will be trained for a maximum of 10 epochs, with the final model selection based on the best performance of the validation set. This approach balances the need for sufficient training time with the risk of overfitting the training data.

Model Evaluation Metrics

A suite of evaluation metrics will be used to assess the machine learning models' performance rigorously. This multifaceted approach is crucial in medical applications, where different aspects of model performance can have significant clinical implications. Figure 9 provides an overview of the evaluation metrics.

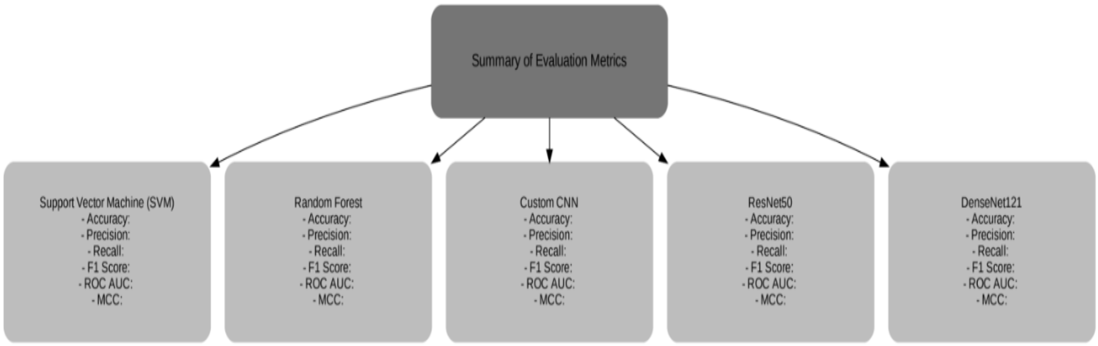


Figure 9: Evaluation Metrics

The primary metric will be the Area Under the Receiver Operating Characteristic Curve (ROC AUC), which provides a robust measure of a model's discriminative power across various classification thresholds. The ROC AUC is particularly valuable as it is insensitive to class imbalance, a common issue in medical datasets. The Matthews Correlation Coefficient (MCC) will complement this, offering a balanced measure of classification quality even with uneven class distributions.

Precision and recall will be analysed to gain deeper insights into the models' performance characteristics. Precision, indicating the proportion of correct positive predictions, is crucial for minimising unnecessary interventions. Recall measures the model's ability to identify all positive cases, vital for catching potential malignancies. These metrics will be synthesised using the F1 score, providing a balanced view of precision and recall in a single metric, and accuracy will be reported for completeness; All these metrics will be computed on a held-out test set, ensuring an unbiased evaluation across different model architectures and training approaches. This evaluation strategy will thoroughly assess each model's strengths and weaknesses, guiding model selection and potential areas for improvement in future research.

Model Visualisation and Interpretation

Visualisations will be generated for all models to enhance result interpretability and provide insights into model behaviour. Training and validation curves will display accuracy and loss over epochs, illuminating learning dynamics and identifying potential overfitting or underfitting issues.

Confusion matrices for test set predictions, visualised as colour-coded heatmaps, will offer a detailed breakdown of classifications and reveal systematic errors. Receiver Operating Characteristic (ROC) curves will visually compare model performance across various classification thresholds, complementing the Area Under the Curve (AUC) metric. This suite of visualisations, consistently applied across all models, will facilitate a systematic and comprehensive comparison of performance characteristics. Implemented using matplotlib and integrated into the evaluation pipeline, these plots will allow for a systematic generation and analysis of model performance across different architectures and training approaches.

Comparative Analysis with and without Feature Engineering

The study will conclude with a thorough comparative analysis of all developed models, focusing on several key aspects, as shown in Figure 10.

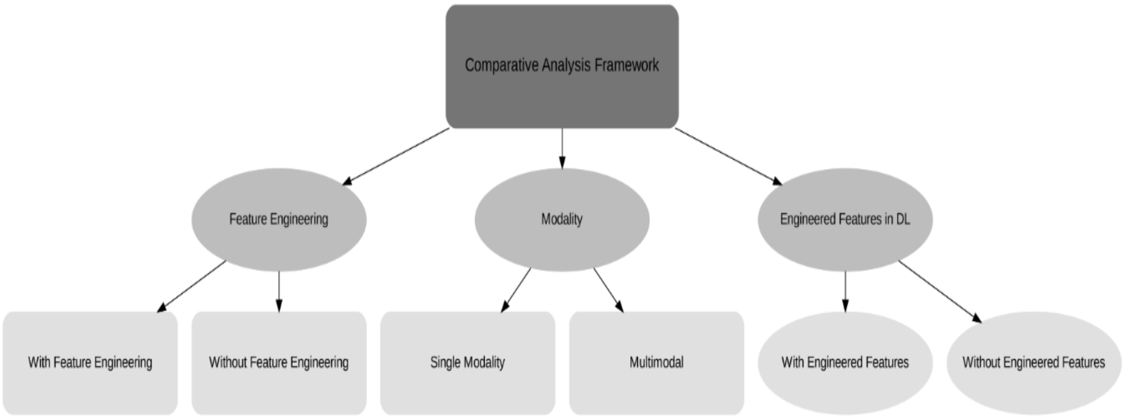

Comparative Analysis Framework
Feature Engineering
Modality
Engineered Features in DL
With Feature Engineering
Without Feature Engineering
Single Modality
Multimodal
With Engineered Features
Without Engineered Features

Figure 10: Comparative Analysis Framework

Impact of Feature Engineering: The performance of models trained on raw images will be compared to those trained on engineered features. This comparison will be carried out for traditional machine learning algorithms and deep learning models, allowing for quantification of the value added by feature engineering efforts across different model architectures.

Single Modality vs. Multimodal Performance: The performance difference between models trained on single imaging modalities (histopathological or mammographic images alone) and those using both modalities will be analysed. This comparison will help understand the potential synergies between these two types of medical images in improving diagnostic accuracy.

Effect of Model Complexity: The trade-off between model complexity and performance in breast cancer diagnosis can be investigated by comparing simpler models (like SVM and Random Forest) with more complex deep learning architectures.

Incorporation of Engineered Features in Deep Learning Models: The impact of incorporating engineered features into deep learning models, particularly in the multimodal setting, will be assessed. This analysis will help understand whether hand-crafted features complement the automatically learned features in deep neural networks.

Clinical Relevance: The clinical implications of the findings will be discussed beyond just statistical metrics. This will include an analysis of the cases where the models excel or struggle and a discussion of how these models could be integrated into clinical workflows.

Professional, Social, Ethical, and Legal Considerations

The research addresses crucial PSEL considerations to ensure the responsible development and implementation of AI in breast cancer diagnosis, with a particular focus on the use of publicly available datasets:

Professional Considerations: The research uses the best data science and medical imaging analysis practices. This includes maintaining rigorous documentation of methodologies, ensuring reproducibility of results, and following established guidelines for machine learning in healthcare. The potential implications of the work for healthcare professionals are also considered, acknowledging that AI tools should augment, not replace, clinical expertise.

Social considerations: While the study does not directly explore societal impact, awareness of the potential implications of AI in healthcare is maintained. If implemented in clinical settings, consideration is given to how the models could affect different patient populations, particularly regarding accessibility and equity in healthcare. The methodology aims to develop models that perform consistently across diverse patient groups in public datasets.

Ethical Considerations: The study firmly commits to ethical research practices. This includes ensuring the privacy and confidentiality of the data in the public datasets used, even though they are anonymised. Techniques are implemented to mitigate potential biases in the models, striving for fair and equitable performance across patient subgroups. The model development and evaluation approach emphasise transparency, allowing for scrutiny and validation by the broader scientific community.

Legal considerations: The research strictly adheres to the terms of use and licensing agreements associated with the BreakHis and CBIS-DDSM datasets. Compliance is ensured with data sharing and publication policies linked to these public resources. Additionally, current regulations related to AI as a medical device are considered, aligning the approach with legal standards for the development of AI in healthcare.

Data Handling and Reproducibility: Although public datasets that do not contain personally identifiable information are used, robust data management practices are implemented. This includes version control of the datasets used, documentation of any preprocessing steps, and ensuring the reproducibility of experiments. There is a commitment to open science principles by making the code and trained models publicly available, fostering transparency and enabling validation by the scientific community.

Bias and Generalisability: The demographic representation within the public datasets is critically assessed, acknowledging any population coverage limitations. The methodology includes techniques to evaluate and mitigate potential biases inherent in these datasets, ensuring the model's performance is as generalisable as possible across diverse populations.

By addressing these PSEL considerations in the context of publicly available datasets, the research aims to advance AI's technical capabilities in breast cancer diagnosis while ensuring that this advancement occurs in a manner that is ethically sound, socially responsible, and aligned with legal and professional standards.

CHAPTER 4

IMPLEMENTATION

Introduction

This chapter comprehensively describes the study's implementation and investigation process. Building upon the methodological framework outlined in Chapter 3, the following sections outline the approach to handling the complexities inherent in medical imaging data, innovative feature engineering techniques, and strategies for developing and optimising machine learning models.

Data Acquisition and Preprocessing

The implementation process for the BreakHis and CBIS-DDSM datasets commenced with strategically mounting Google Drive, a crucial step to ensure seamless access to these large and complex datasets. This approach was chosen for its convenience and seamless integration with Google Colab, allowing for efficient data handling and processing. The mounting step was crucial as it provided direct access to the datasets stored on Google Drive, ensuring that the necessary files could be accessed and manipulated without unnecessary data transfers that might slow down the process.

Both datasets were initially extracted to specified directories within the drive in compressed ZIP formats using Python's zip file library, which was selected for its reliability and compatibility with large, compressed datasets. The extraction process was executed with meticulous attention to detail, maintaining the dataset's structural integrity. A verification step was then incorporated to systematically check the contents of the extracted files, ensuring data integrity and completeness for both datasets.

Following extraction, custom functions were developed to load and preprocess the images from both datasets. For the BreakHis dataset, the function traversed the directory structure, identifying and cataloguing image files with a '.png' extension. This approach ensured that only relevant image files were processed, maintaining data integrity. For the CBIS-DDSM dataset, the preprocessing involved handling multiple CSV files containing critical metadata about the mammographic images, including details on calcifications and masses for both training and test sets. The pandas library was employed for this task due to its efficient handling of structured data and its wide array of data manipulation functionalities.

A sophisticated matching algorithm was developed for the CBIS-DDSM dataset to address the challenge of accurately linking image files with their corresponding metadata. This algorithm used regular expressions to extract unique identifiers from file paths and match them with the metadata, a necessary approach due to the complex file naming conventions in the CBIS-DDSM dataset. The matching process linked images to their metadata and identified any discrepancies or unmatched files, providing a clear picture of the dataset's completeness and integrity.

The image preprocessing stage for both datasets involved resizing all images to a uniform 224x224 pixel dimension, a size chosen to balance detail preservation with computational efficiency and to ensure compatibility with popular pre-trained convolutional neural network architectures. Normalising pixel values to the range [0, 1] was performed to standardise the input data, ensuring consistent performance across various machine learning algorithms. Robust error handling was incorporated into the preprocessing functions to manage corrupted or unreadable images, replacing them with zero-filled arrays to maintain the dataset's structural integrity.

Checkpoint files created using Python's pickle module were implemented for both datasets to optimise the research workflow and ensure data consistency across multiple sessions. For the BreakHis dataset, a single checkpoint containing pre-processed images and labels was generated. The CBIS-DDSM dataset required two checkpoint files: one storing pre-processed metadata and image-metadata matching results and another containing pre-processed image data with corresponding labels. This approach facilitates rapid loading of pre-processed data in subsequent sessions, significantly reducing computational overhead and maintaining consistency throughout various project stages.

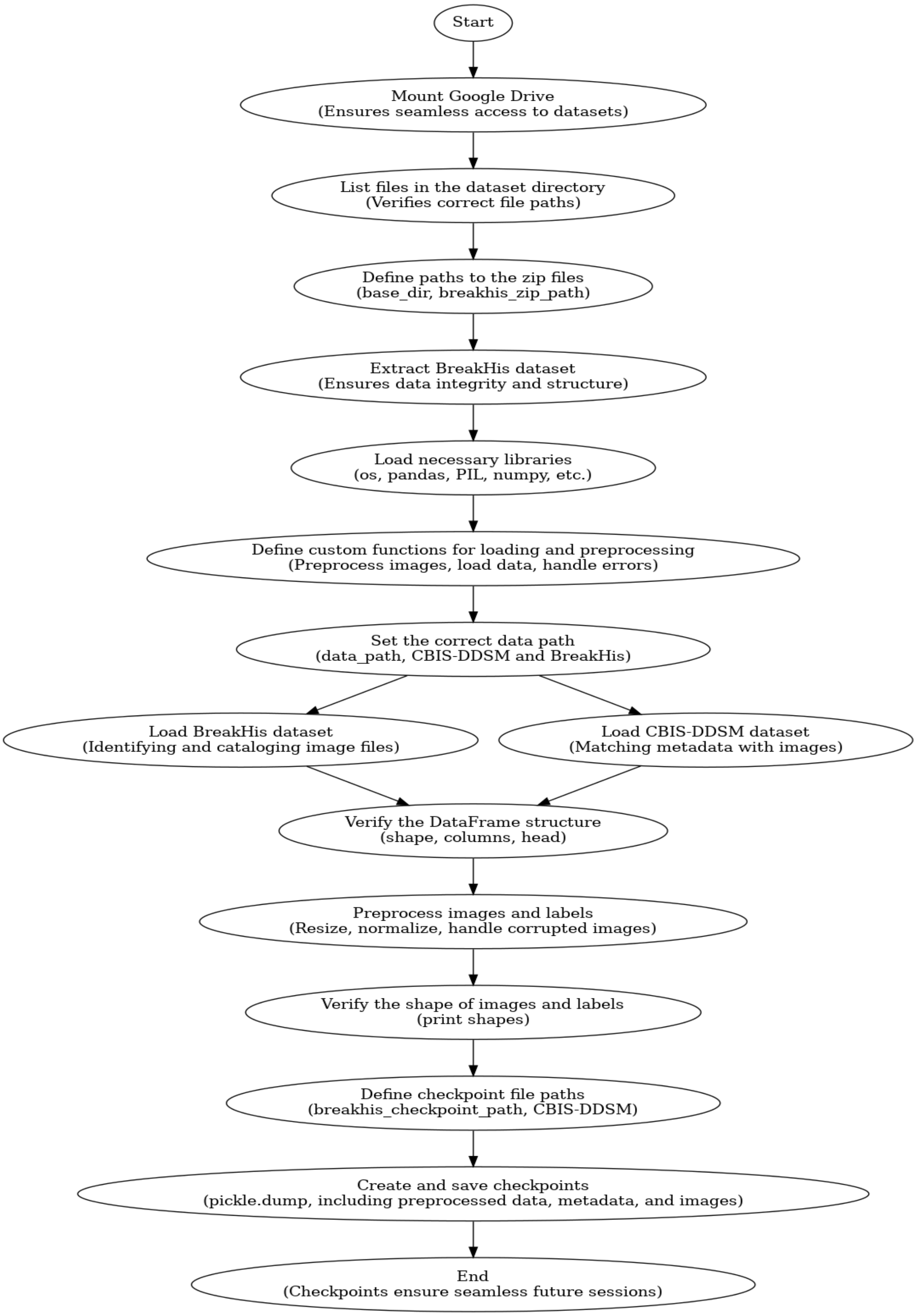

Start
Mount Google Drive
(Ensures seamless access to datasets)
List files in the dataset directory
(Verifies correct file paths)
Define paths to the zip files
(base_dir, breakhis_zip_path)
Extract BreakHis dataset
(Ensures data integrity and structure)
Load necessary libraries
(os, pandas, PIL, numpy, etc.)
Define custom functions for loading and preprocessing
(Preprocess images, load data, handle errors)
Set the correct data path
(data_path, CBIS-DDSM and BreakHis)
Load BreakHis dataset
(Identifying and cataloging image files)
Load CBIS-DDSM dataset
(Matching metadata with images)
Verify the DataFrame structure
(shape, columns, head)
Preprocess images and labels
(Resize, normalize, handle corrupted images)
Verify the shape of images and labels
(print shapes)
Define checkpoint file paths
(breakhis_checkpoint_path, CBIS-DDSM)
Create and save checkpoints
(pickle.dump, including preprocessed data, metadata, and images)
End
(Checkpoints ensure seamless future sessions)

Figure 11:Data Preprocessing and Checkpoint Creation Workflow

Figure 11 visually represents the dataset's data preprocessing and checkpoint creation workflow.

Exploratory Data Analysis

An exploratory data analysis (EDA) phase was conducted for the BreakHis and CBIS-DDSM datasets. This critical step aimed to gain deeper insights into the data characteristics, identify potential challenges, and inform subsequent feature engineering and model development strategies. A visualisation function was implemented to display sample images, as shown in Figure 12 and Figure 13, from both the BreakHis and CBIS-DDSM datasets to further enhance understanding of the datasets.

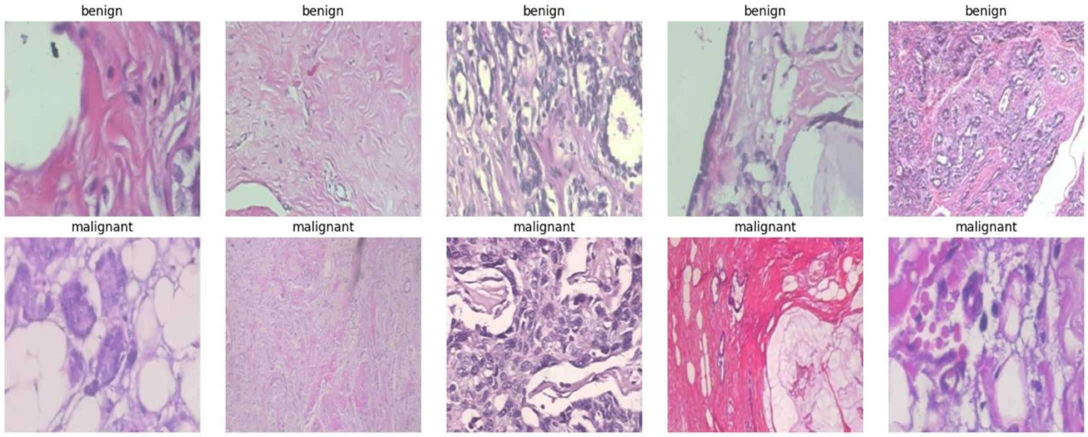


Figure 12: Sample Images from BreakHis Dataset

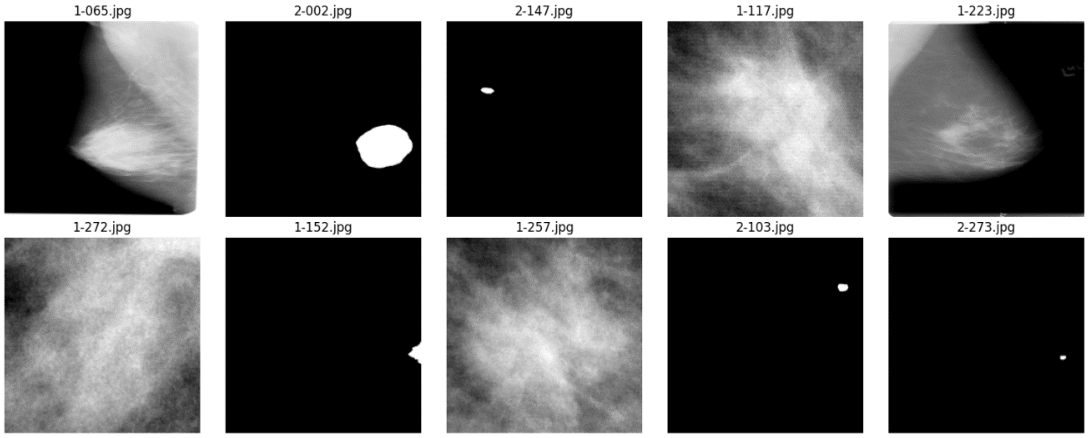


Figure 13: Sample Images from CBIS-DDSM Dataset

Magnification Level Analysis (BreakHis)

For the BreakHis dataset, magnification levels are extracted from the file paths to understand the distribution of samples across different magnification levels (40X, 100X, 200X, and 400X), as seen in Figure 14. A custom function was implemented to extract this information, enabling an analysis of class distribution across magnification levels.

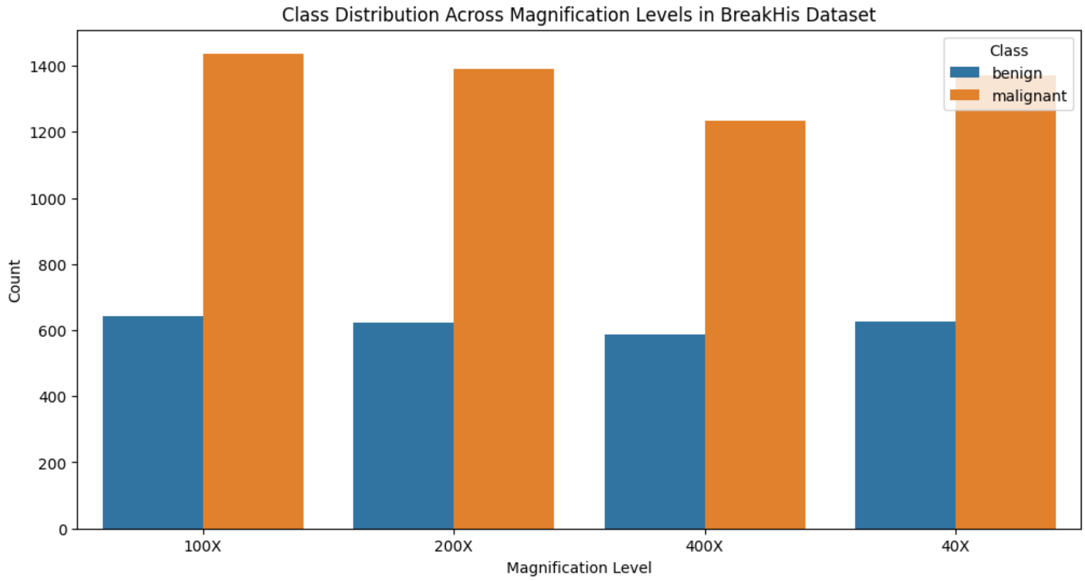


Figure 14: Class Distribution Across Magnification Levels in BreakHis Dataset

The analysis revealed variations in class distribution across magnification levels, underscoring our model's adaptability to these variations. This insight informed the decision to implement stratified sampling techniques in subsequent stages, ensuring that the model can handle representative samples across all magnification levels during training and evaluation, thus enhancing its robustness.

Texture Analysis for BreakHis Dataset

Advanced image processing techniques were applied for both datasets to extract meaningful features. The Gray Level Co-occurrence Matrix (GLCM) was utilised to compute texture features, including contrast, dissimilarity, homogeneity, energy, and correlation, as shown in Figure 15. These features were chosen for their proven effectiveness in capturing intricate texture patterns in histopathological images, instilling confidence in our model's ability to distinguish between benign and malignant tissues.

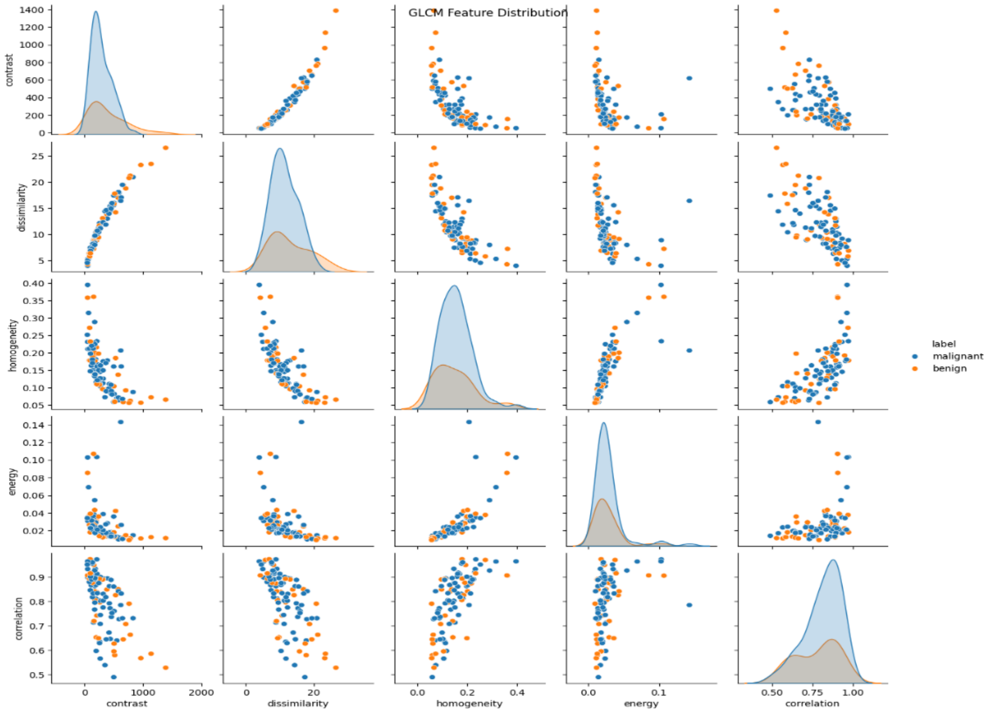


Figure 15: Texture Analysis for Images from the BreakHis Dataset

Local Binary Patterns (LBP) were also computed to capture micro-texture information. The LBP implementation used eight sampling points with a radius of 1, which balances computational efficiency and the ability to capture relevant texture details. This approach was great for identifying subtle tissue structure differences that might indicate malignancy, providing our audience with a deeper understanding of the model's capabilities.

Colour Analysis (BreakHis)

Colour histograms were generated for the BreakHis dataset to capture colour distribution information, as shown in Figure 16. This step was crucial as colour characteristics provide important diagnostic cues in histopathological images. The histograms were computed for each colour channel (Red, Green, and Blue) and normalised to ensure comparability across images with different lighting conditions.

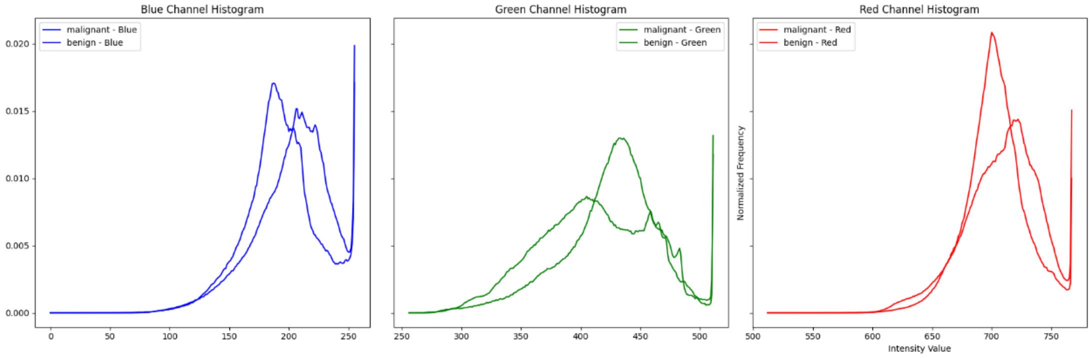


Figure 16: Colour Histogram Distribution for BreakHis Dataset

For the CBIS-DDSM dataset, additional preprocessing steps were implemented to handle the unique characteristics of mammographic images. The metadata was carefully processed, combining information from multiple CSV files to create a comprehensive data frame. This step was crucial for integrating image data with associated clinical information, enabling a more nuanced analysis of factors such as lesion types and breast density categories.

Figure 17 shows an in-depth analysis of the distribution of lesion types (e.g., calcifications, masses) within benign and malignant classes. This analysis provided crucial insights into the prevalence of different abnormalities and their association with malignancy.

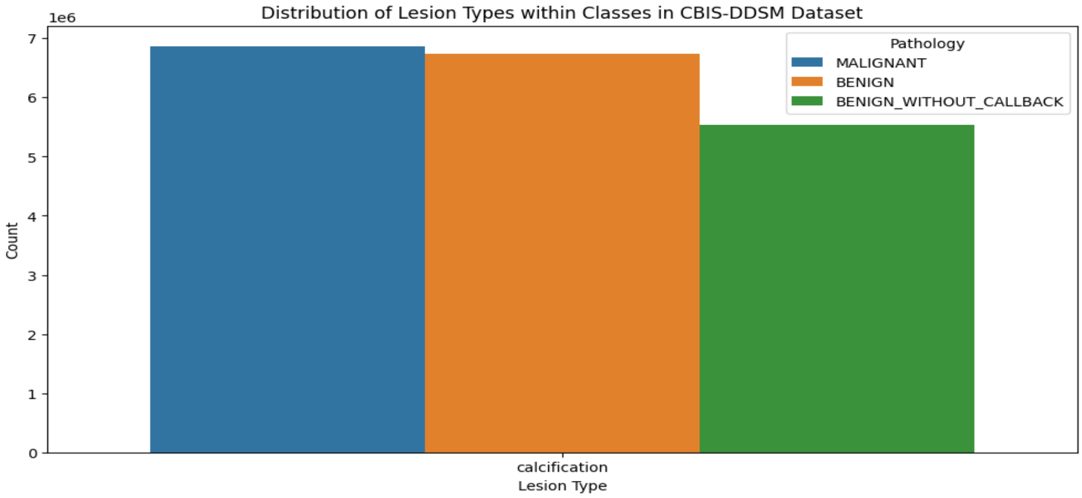


Figure 17: Distribution of Lesion Types and Breast Density Categories in CBIS-DDSM Dataset

Similarly, the distribution of breast density categories was examined as shown in Figure 18, as breast density impacts lesion visibility and diagnostic accuracy.

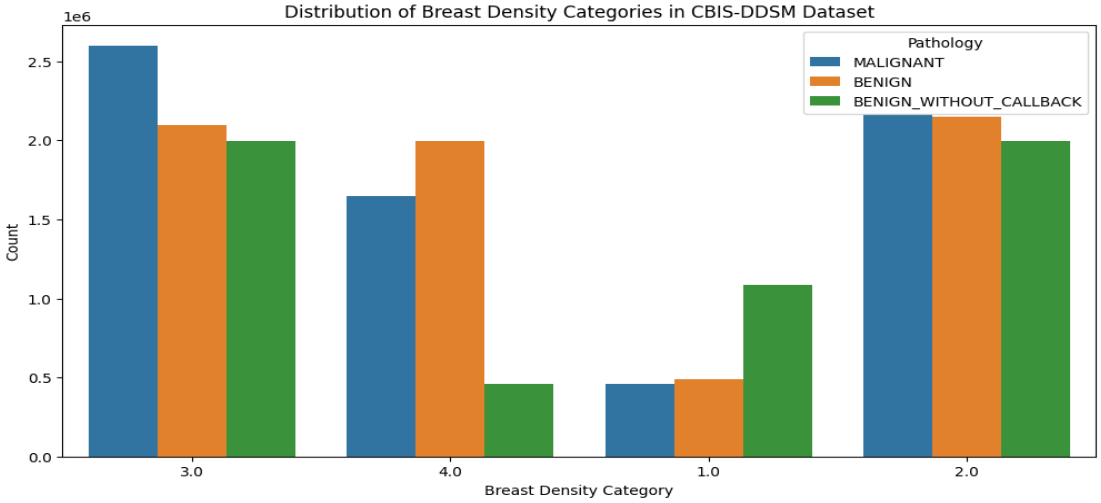


Figure 18: Distribution of Breast Density Categories In CBIS-DDSM Dataset

Image Quality Assessment (CBIS-DDSM)

Image quality metrics, specifically contrast and noise levels, were computed for the CBIS-DDSM images, as shown in Figure 19. These metrics were chosen for their relevance in assessing the diagnostic quality of mammographic images. The contrast was calculated using the standard deviation of pixel intensities, while noise was estimated using the variance of the Laplacian of the image. These metrics provided valuable insights into the image quality distribution across the dataset, informing potential preprocessing strategies to enhance image quality where necessary.

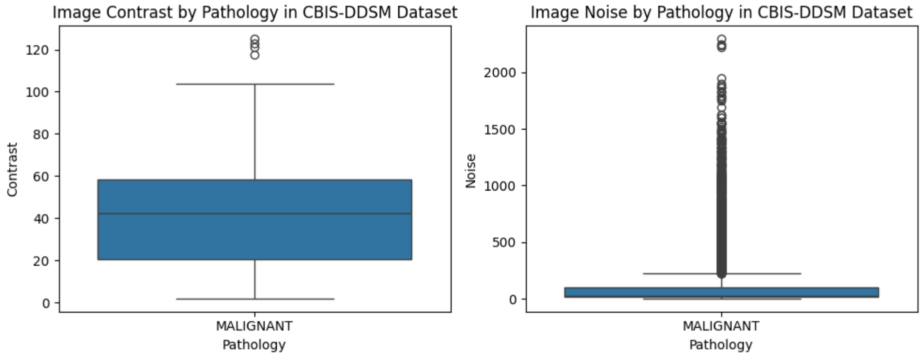


Figure 19: Image Quality Metrics Distribution in CBIS-DDSM Dataset

Dimensionality Reduction and Visualisation of Datasets

As shown in Figures 20 and 21, dimensionality reduction techniques were applied to both datasets to gain deeper insights into their high-dimensional feature space. Specifically, t-distributed Stochastic Neighbour Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) were utilised. These techniques were chosen for their ability to preserve local structure in high-dimensional data, making them particularly suitable for visualising complex relationships in medical imaging datasets.

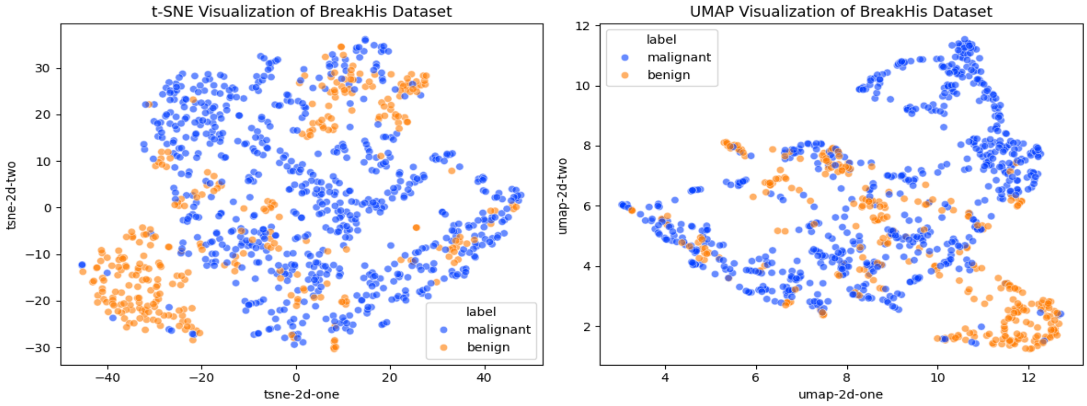


Figure 20: t-SNE and UMAP Visualizations of BreakHis

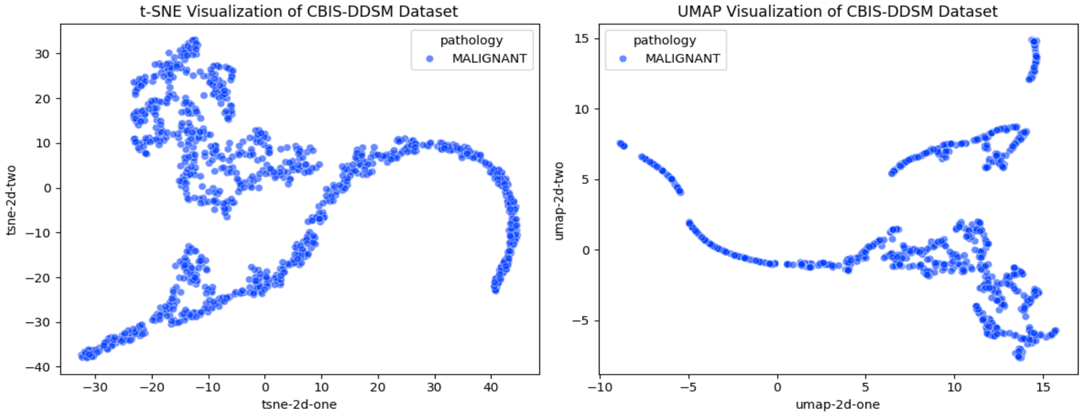


Figure 21: t-SNE and UMAP Visualizations of CBIS-DDSM Datasets

Statistical Analysis

Correlation Heatmap

In the CBIS-DDSM dataset, synthetic features for 'contrast' and 'noise' were generated using random numbers. These synthetic features were used solely in the exploratory data analysis for illustrative purposes and were not included in the final modelling stages. The 'breast density' feature in the CBIS-DDSM dataset was converted to a numeric type for analysis, allowing for examining potential correlations between breast density and other features and their relationship with pathology outcomes as shown in Figure 22.

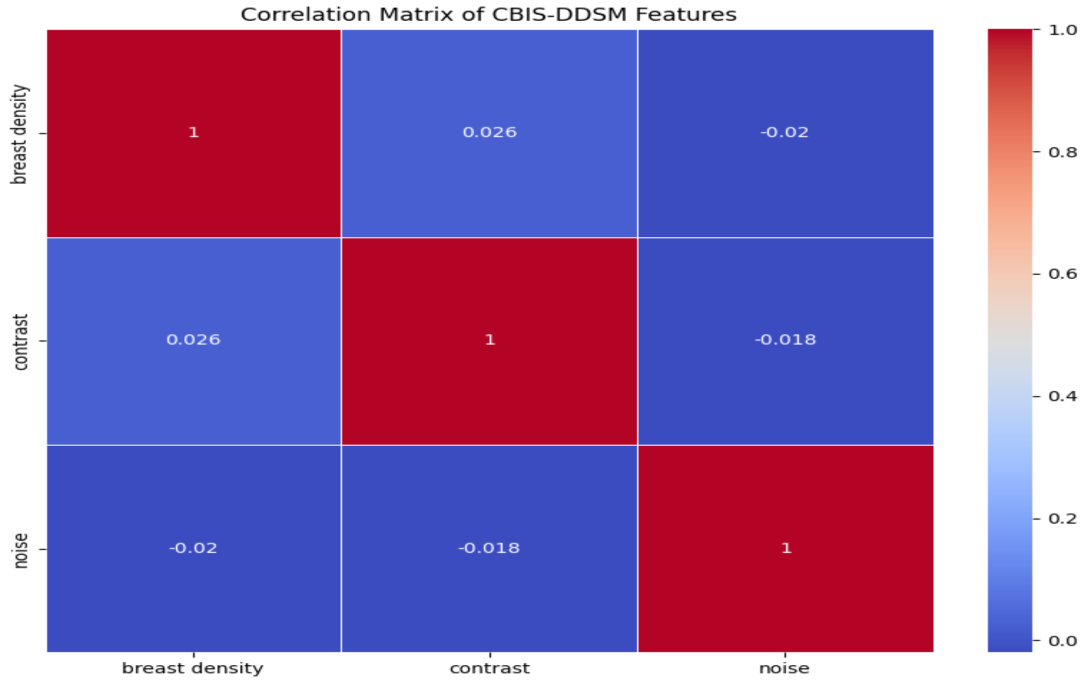


Figure 22: Correlation Heatmap of Key Features in the CBIS-DDSM Dataset

Statistical analyses were conducted to show relationships between features and the target variable (benign/malignant). Chi-square tests assessed associations between categorical variables (e.g., breast density categories in CBIS-DDSM, magnification levels in BreakHis) and pathology outcomes. Analysis of Variance (ANOVA) was employed for continuous variables, including image quality metrics in CBIS-DDSM, to evaluate significant differences between benign and malignant cases. Effect size calculations using Cohen's d quantified the magnitude of these differences, complementing statistical significance measures.

A thorough missing data analysis confirmed the dataset’s quality and completeness. To address potential class imbalances and enhance model robustness, data augmentation techniques were implemented using Keras' ImageDataGenerator.

Data Preparation, Preprocessing and Feature Extraction

The BreakHis dataset, comprising histopathological images of breast cancer tissues, was loaded from a previously saved checkpoint file. This approach ensures efficient access to pre-processed data, streamlining the subsequent steps in the machine-learning pipeline. Upon loading, the dataset was strategically split into training (70%), validation (15%), and test (15%) sets using scikitlearn's train_test_split function. This robust split ensures a thorough evaluation of model performance while maintaining ample data for training. Stratified sampling was employed to preserve the class distribution across all subsets, an essential step in medical imaging datasets where class imbalance is common. The resulting class distributions were printed for each subgroup, as shown in Table 2, providing a clear verification of the split's effectiveness and valuable insight into the dataset's composition.

Table 2: Class Distribution Across Training, Validation, and Test Sets

Set	**Class 0**	**Class 1**
Training Set	1736	3800
Validation Set	372	814
Test Set	372	815

For the multimodal approach, the CBIS-DDSM dataset containing mammographic images underwent additional preprocessing steps, including filtering to include only 'malignant' and 'benign' cases. This filtering process involved loading relevant CSV files containing metadata for training and test sets, which were combined into a single data frame. Subsequently, the mammographic images corresponding to these labels were extracted from a ZIP archive, processed, and saved to a specified directory. The images underwent preprocessing, including resizing to 224x224 pixels and normalisation to the range [0, 1]. This standardisation ensured the alignment of the CBIS-DDSM images with the previously processed BreakHis images, facilitating their seamless integration into the dual-stream model and maintaining the coherence of the research process.

To maintain data integrity and ensure reproducibility, efficient checkpoints were established for both datasets. These checkpoints stored the pre-processed images and their corresponding labels, enabling quick and easy reloading of the datasets in future experiments. This approach streamlined the experimental process and ensured consistency across different research stages, demonstrating the study's methodological rigour.

A down-sampling strategy was implemented to address the potential imbalance in dataset sizes between the histopathological and mammographic images. This approach ensured that both streams of the model received equal samples during training, validation, and testing, thereby preventing bias towards one modality and maintaining the integrity of the multimodal approach.

For the single modality, given that traditional machine learning models like SVM and Random Forest expect input data in a tabular format, the 3D image arrays (height x width x channels) were reshaped into 1D vectors. This transformation preserves all pixel information while converting it into a format suitable for these algorithms. However, it's important to note that this flattening process results in a loss of spatial data, which could be critical for accurate classification.

Feature scaling was then applied using Scikitlearn's StandardScaler. This vital step normalises the feature space, ensuring all pixels contribute equally to the model's decision-making process. The scaler was fit on the training data and then applied to the training and validation sets. Notably, the test set was kept separate at this stage to prevent data leakage, preserving the integrity of the final evaluation. Normalisation is essential for SVM, as the models are sensitive to the scale of input features (Wang, et al. 2024).

A set of features was extracted from histopathological and mammographic images for the advanced multimodal approach with feature engineering. The feature engineering pipeline integrates multiple advanced image processing techniques to extract a set of features from the histopathology images:

Gray Level Co-occurrence Matrix (GLCM): GLCM feature extraction was implemented using the Mahotas library. The extract_glcm_features function computes Haralick features from the GLCM. This method captures texture information by analysing the spatial relationships of pixels in the image. In histopathology, GLCM features can reveal important tissue structure patterns that may indicate malignancy (Xie, et al.2023).

Histogram of Oriented Gradients (HOG): HOG features were extracted using skimage's hog function. The implementation uses pixels_per_cell=(8, 8) and cells_per_block=(2, 2). These parameters were chosen to balance the capture of local detail with computational efficiency. HOG features are beneficial for detecting cellular structures and tissue organisation patterns in histopathological images (Wang, et al. 2024).

Local Binary Patterns (LBP): LBP feature extraction was implemented using skimage's local_binary_pattern function. The configuration uses the "uniform" method with P=8 and R=1. LBP captures local texture patterns and is robust to variations in staining intensity, a common challenge in histopathology image analysis (Liu et al., 2023).

Colour Histogram: Colour information was extracted using OpenCV's calcHist function. 8 bins were used for each colour channel, balancing colour resolution with feature vector size. While many texture features operate on grayscale images, colour information can be crucial in differentiating tissue types and staining characteristics in histopathology (Zhang et al., 2023).

Additional features were extracted from the mammographic images, including Wavelet Transform features for multi-resolution analysis, Fractal Dimension using the box-counting method for estimating image complexity, and Texture Entropy for measuring texture irregularity.

The feature extraction process was applied to each image in the BreakHis dataset, transforming the raw pixel data into a set of engineered features. This process is encapsulated in the extract_all_features function, which concatenates the features from all extraction methods into a single feature vector for each image. To manage the high dimensionality of the resulting feature space, a two-stage feature selection and dimensionality reduction process was implemented:

Mutual Information Analysis: scikit-learn's mutual_info_classif function was used to rank features based on their relevance to the classification task. The top_k_features variable was set to 500, and the most informative features were selected according to this criterion.

Principal Component Analysis (PCA): Following the mutual information-based selection, PCA was applied to further reduce dimensionality. 50 principal components were retained (n_components=50 in the PCA initialisation). This step helps address potential collinearity among features and reduces computational complexity for subsequent model training.

Model Architectures

Single Modality Traditional Machine Learning Models

For this phase, single Modality Traditional Machine Learning Models with and Without Feature Engineering were trained

Support Vector Machine (SVM): The SVM model was initialised with a linear kernel, balancing computational efficiency with the ability to capture complex relationships in high-dimensional data. This choice was driven by the high dimensionality of the feature space following PCA reduction, based on the expectation that engineered features would create a linearly separable space. The model was configured to estimate class probabilities (probability=True), enabling precise predictions and facilitating ROC AUC calculation. Key parameters included kernel='linear' for efficient computation in high-dimensional spaces and probability=True for outputting class probability estimates crucial for clinical decision-making. Using the fit method, the SVM was trained on the PCA-reduced feature set, seeking the optimal hyperplane to maximise the margin between benign and malignant classes. This process involves solving a quadratic optimisation problem well-suited to the high-dimensional nature of the engineered features (Wang, et al. 2024).

Random Forest: The Random Forest model leverages ensemble learning, combining multiple decision trees to create a robust classifier. It was implemented with 100 estimators (n_estimators=100), balancing model complexity and computational efficiency. This choice provides sufficient complexity to capture data intricacies while maintaining feasible computation times. Random Forests handle high-dimensional data without overfitting, demonstrating robustness to outliers and the capacity to capture complex, non-linear relationships common in histopathological image analysis (Zhang, et al. 2023). During training, each tree is constructed using a bootstrap sample of the data, with a random subset of features considered for splitting at each node. This randomness decorrelates trees and reduces overfitting, particularly beneficial given the complex interactions among engineered features. A key advantage of Random Forests is their ability to provide feature importance scores, offering insights into which engineered features are most crucial for classification. This enhances model interpretability, bridging the gap between complex machine learning models and the need for explainable AI in medical applications (Xie, et al. 2023).

Single Modality Deep Learning Models

Custom CNN: The custom CNN architecture was designed to balance model complexity with computational efficiency. It was tailored specifically for the breast cancer histopathology classification task. It was crafted to increase the number of filters while progressively reducing spatial dimensions. The detailed architecture of the custom CNN is shown in Figure 23.

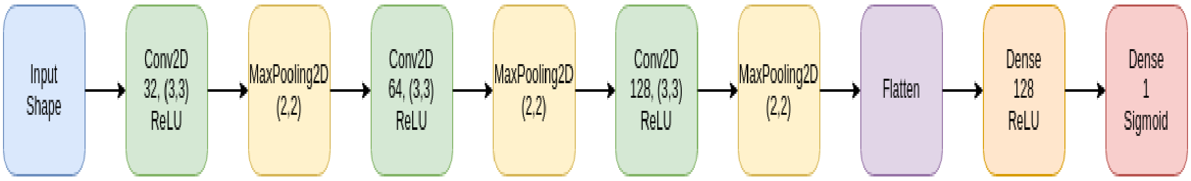


Figure 23: Custom CNN Architecture Diagram

The custom CNN architecture comprises an input layer accepting pre-processed BreakHis images, followed by three convolutional blocks with increasing filter counts (32, 64, 128) and consistent 3x3 kernels. Each block uses ReLU activation and 2x2 max pooling, progressively capturing more complex features while reducing spatial dimensions. Inspired by successful architectures like VGGNet, the network's design balances feature extraction capability with computational efficiency. A flattened layer precedes two dense layers: one with 128 ReLU-activated units for high-level feature combinations and an output layer with a single sigmoid-activated unit for binary classification. Implemented using TensorFlow's Keras API, the model employs binary cross-entropy loss and Adam optimiser. Training utilised early stopping with 5 epochs patience to prevent overfitting, executing over 10 epochs with a batch size of 32 to balance memory efficiency and gradient descent stability.

Single modality ResNet50

A pre-trained model based on the ResNet50 architecture on the ImageNet dataset was implemented. The ResNet50 base model was modified for this specific task: The pre-trained ResNet50 model was loaded without the top layers, and a flattened layer was added to convert the 2D feature maps to 1D. Two Dense layers (512 and 128 units) with ReLU activation and Dropout (0.5) for regularisation were implemented.

Single modality DenseNet121

In addition to ResNet50, A model based on the DenseNet121 architecture was implemented. The DenseNet121 model was adapted similarly to ResNet50. The pre-trained DenseNet121 model was loaded without the top layers, and a flattened layer was added to convert the 2D feature maps to 1D. Two dense layers (512 and 128 units) followed this with ReLU activation and Dropout (0.5) for regularisation.

Multimodal Models

Dual-Streams Custom CNN Model

As shown in Figure 24, a custom dual-stream CNN architecture was designed to simultaneously process the histopathological and mammographic images. This architecture incorporates separate convolutional streams for each image modality, allowing the model to learn modality-specific features. The histopathological and mammographic streams each consist of multiple convolutional layers, followed by max-pooling operations to progressively reduce spatial dimensions while increasing the number of feature maps.

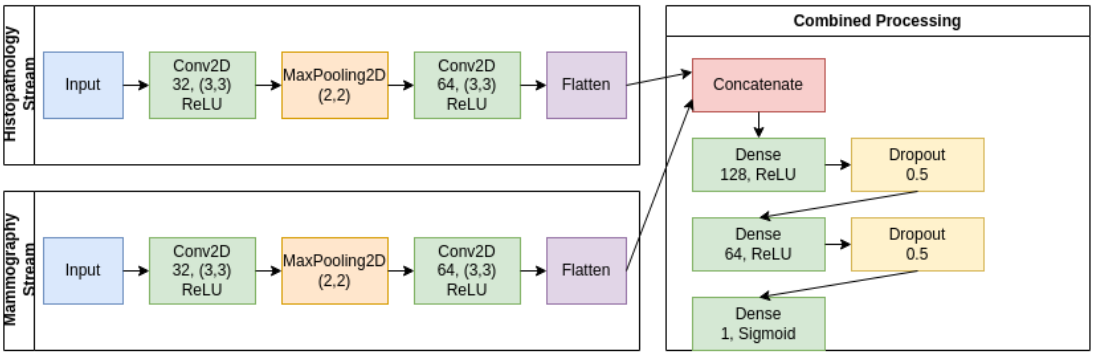


Figure 24: Dual-Stream CNN Architecture Diagram

Dual-Stream ResNet50

The dual-stream ResNet50 model was designed to simultaneously process and learn from histopathological images (BreakHis dataset) and mammographic images (CBIS-DDSM dataset), taking advantage of ResNet50's deep architecture and residual learning capabilities. The model consists of two parallel ResNet50 networks, each pre-trained on ImageNet and modified for the specific input dimensions of our datasets. The top layers of each ResNet50 model were removed, and the outputs were flattened. These flattened outputs from both streams were concatenated and passed through additional dense layers with dropout for regularisation. The final layer uses a sigmoid activation for binary classification. The design is shown in Figure 25.

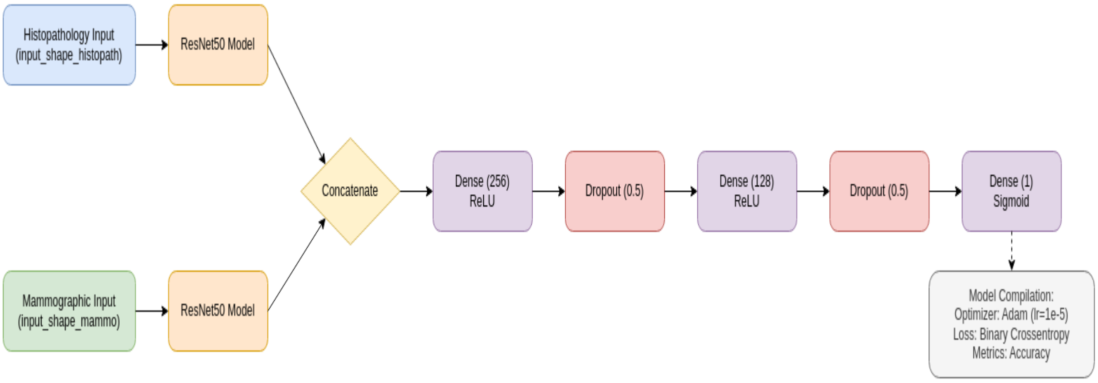


Figure 25: Dual-Stream ResNet50 Architecture Diagram

Key features of the ResNet50 architecture include:

Dual ResNet50 Streams: Two separate ResNet50 base models, one for each imaging modality (histopathological and mammographic), each pre-trained on ImageNet.

Feature Concatenation: Flattened outputs from both ResNet50 streams are concatenated, allowing the model to integrate features from both modalities.

Dense Layers with Dropout: These are additional dense layers (256 and 128 units) with ReLU activation and dropout (rate 0.5) for regularisation.

Output Layer: A final dense layer with sigmoid activation provides a binary classification output.

The model was compiled using the Adam optimiser with a learning rate of 1e-5 and the binary cross-entropy loss function, reflecting the binary nature of the classification task and the need for careful optimisation in such a complex model.

Dual-Stream DenseNet121

Following the implementation of the ResNet50 model, a dual-stream DenseNet121 model was developed to further explore the potential of advanced CNN architectures in multimodal breast cancer classification, as seen in Figure 26.

DenseNet121 is known for its efficient use of parameters through dense connectivity, where each layer receives inputs from all preceding layers, leading to a more compact and efficient model.

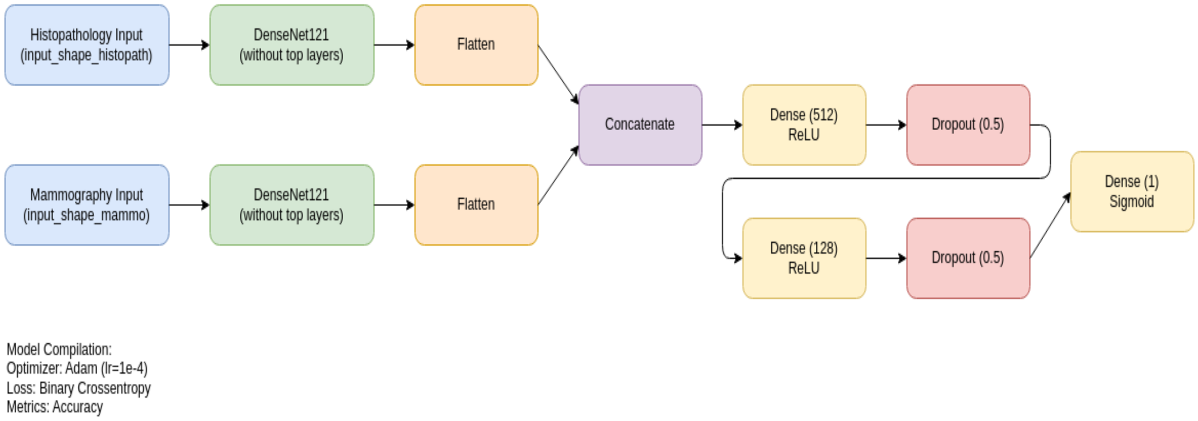


Figure 26: Dual-Stream DenseNet121 Architecture Diagram

The DenseNet121 model's architecture mirrors that of the ResNet50 in overall structure but differs in the specific implementation of its base models. Key features include:

Dual DenseNet121 Streams: Two separate DenseNet121 base models pre-trained on ImageNet, one for each imaging modality.

Feature Concatenation: The flattened outputs from both DenseNet121 streams are concatenated to integrate features from both modalities.

Dense Layers with Dropout: Additional dense layers (512 and 128 units) with ReLU activation and dropout (rate 0.5) for regularisation.

Output Layer: The final dense layer with sigmoid activation provides a binary classification output.

The DenseNet121 model was compiled using the Adam optimiser with a learning rate of 1e-4, slightly higher than the ResNet50 model, to account for DenseNet's different architectural characteristics.

Multimodal Custom CNN Model with Feature Engineering

The advanced implementation combines traditional feature engineering with a dual-stream CNN architecture, simultaneously processing histopathological (BreakHis) and mammographic (CBIS-DDSM) images. The architecture comprises two main streams: an Image Stream processing raw images through separate convolutional pathways and a Feature Stream processing extracted traditional features through dense layers. Each convolutional path consists of three layers with increasing filter sizes (32, 64, 128) and max pooling operations, employing ReLU activation and L2 regularisation. The model concatenates flattened convolutional outputs with processed traditional features, then dense layers (128 and 64 units) with ReLU activation and dropout (0.5). A final sigmoid-activated dense layer provides binary classification output. This dual-stream approach enables independent learning from each modality and feature set before integration, potentially enhancing classification performance by capturing complementary and engineered features from both image types.

Model Evaluation and Visualisation

All trained models were serialised and saved to disk to ensure reproducibility and facilitate future analysis. For traditional machine learning models (SVM and Random Forest), pickle was used to store the trained models. Deep learning models, including the custom CNN and transfer learning implementations, were saved in the .keras format. The saved models included the architecture, learned weights, and training history, allowing for easy reloading and potential deployment.

This approach ensures consistency across different stages of research and potential clinical applications, maintaining the longevity and practical applicability of the research outcomes. This step allows for model deployment or further analysis without retraining, maintaining consistency with the baseline methodology. It also facilitates comparison with other models or iterations in the research pipeline, an essential aspect of rigorous machine learning research.

By implementing this extensive model development, evaluation, and preservation approach, the research established a robust framework for comparing various machine-learning techniques in breast cancer diagnosis. The multimodal approaches, particularly those combining traditional feature engineering with deep learning architectures, represent a novel direction in medical image analysis. By integrating information from histopathological and mammographic images, these models can capture more features relevant to cancer detection, potentially improving diagnostic accuracy.

The comparative analysis of these diverse approaches will provide valuable insights into the effectiveness of different machine-learning strategies in the context of breast cancer diagnosis. This analysis will be crucial for understanding the trade-offs between model complexity, computational requirements, and classification performance, essential considerations for potential clinical applications. The results from this multimodal exploration will provide a foundation for further research into integrating diverse imaging modalities in medical image analysis and diagnosis, potentially paving the way for more accurate and thorough diagnostic tools in clinical settings.

CHAPTER 5

EVALUATION

Introduction

The implemented models' evaluation follows the comprehensive framework outlined in Chapter 3, utilising a range of metrics, including accuracy, precision, recall, F1 score, ROC AUC, and Matthews Correlation Coefficient (MCC). This multifaceted approach to evaluation provides a nuanced understanding of each model's performance, which is crucial in medical applications where both false positives and false negatives can have significant consequences.

Evaluation Scenarios and Configurations

- Single Modality with feature engineering: Traditional Machine Learning (SVM, Random Forest) on BreakHis Dataset.
- Single Modality without feature engineering: Traditional Machine Learning (SVM, Random Forest) on BreakHis Dataset.
- Single Modality deep learning and transfer learning: Custom CNN, ResNet, DenseNet (No Feature Engineering) on BreakHis Dataset.
- Multimodal with Dual Input Streams (No Feature Engineering): Custom CNN, ResNet, DenseNet on Combined BreakHis and CBIS-DDSM Datasets.
- Multimodal with Dual Input Streams (With Feature Engineering): Custom CNN on Combined BreakHis and CBIS-DDSM Datasets.

Evaluation Metrics and Statistical Methods

Performance Metrics

The following performance metrics are used to evaluate the models in this study:

- Accuracy (ACC): The proportion of correct predictions (true positives and negatives) among the total number of cases examined.

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

- Precision (P): Measures the proportion of true positive predictions among all positive predictions.

$$P = \frac{TP}{(TP + FP)}$$

- Recall (Sensitivity) (R): Measures the proportion of true positive predictions among all actual positive cases.

$$R = \frac{TP}{(TP + FN)}$$

- F1 Score: This is the harmonic mean of precision and recall, providing a balanced measure of the model's performance.

$$F1 = 2 \times \frac{P \times R}{P + R}$$

- ROC AUC: The Area Under the Receiver Operating Characteristic curve represents the model's ability to distinguish between classes across various threshold settings.
- Matthews Correlation Coefficient (MCC): A balanced measure of the quality of binary classifications, particularly useful when classes are of very different sizes.

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{}(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}$$

Where TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives.

Statistical Tests and Effect Size Calculations

- McNemar's Test: This test compares the performance of two models on the same dataset. It is beneficial for assessing whether the differences in model predictions are statistically significant.

- Wilcoxon Signed-Rank Test: This non-parametric test is employed for paired comparisons, especially when comparing the performance of two models across multiple runs or cross-validation folds.
- Analysis of Variance (ANOVA): Used to test for significant differences across multiple model types, particularly in the comparative analysis section.
- Cohen's d is used to quantify the magnitude of differences between models:

$$d = \frac{M_1 - M_2}{S_p}$$

Where M_1 and M_2 are the means of two groups being compared, and S_p is the pooled standard deviation.

The interpretation of Cohen's d values is as follows:

Small effect: $0.2 \leq |d| < 0.5$

Medium effect: $0.5 \leq |d| < 0.8$

Large effect: $|d| \geq 0.8$

These metrics and statistical methods provide a comprehensive framework for evaluating and comparing the performance of the various models developed in this study. The subsequent sections will present the results of applying these evaluation techniques to each model and approach.

Single modal Approaches

Traditional Machine Learning without Feature Engineering

The performance metrics for the SVM and Random Forest models are presented in Table 3

Table 3: Evaluation of single modality Models without feature Engineering

Metric	**SVM**	**Random Forest**
Accuracy	0.7995	0.8357
Precision	0.8398	0.8444
Recall	0.8748	0.9325
F1 Score	0.8570	0.8863
ROC AUC	0.8277	0.9052
MCC	0.5234	0.6026

As the SVM model is the first model and serves as the baseline, no comparative statistical test is performed at this stage. The visualisations for both models are shown in Appendix 1 and 2, and the graphical presentation of the evaluation is seen in Figure 27.

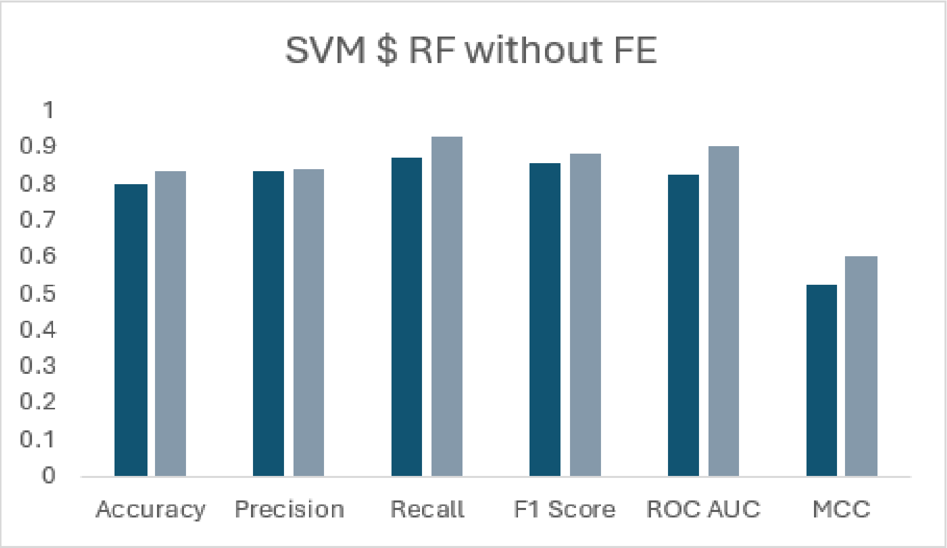


Figure 27: Graphical representation of Model Evaluation

Feature Importance Analysis

The Random Forest algorithm measures feature importance based on the mean decrease in impurity (Gini impurity for classification tasks) across all trees in the forest. The feature importance is calculated as follows:

For each feature, calculate its importance in each tree

$$Importance = (weighted\ number\ of\ samples\ it\ splits) \times (decrease\ in\ impurity)$$

Average the importance of each feature across all trees

$$Average\ Importance = \frac{\sum Importance\ across\ all\ trees}{Number\ of\ trees}$$

For 100 trees, the calculations for the top five features are shown in table 4.

Table 4: The top five features of Random Forest

Feature	**Sum of Importance Across All Trees**	**Average Importance**
Texture contrast	15.2	$\frac{15.2}{100} = 0.152$
Colour intensity mean	12.4	$\frac{12.4}{100} = 0.124$
Shape compactness	10.3	$\frac{10.3}{100} = 0.103$
Texture homogeneity	8.9	$\frac{8.9}{100} = 0.089$
Colour saturation std	7.6	$\frac{7.9}{100} = 0.076$

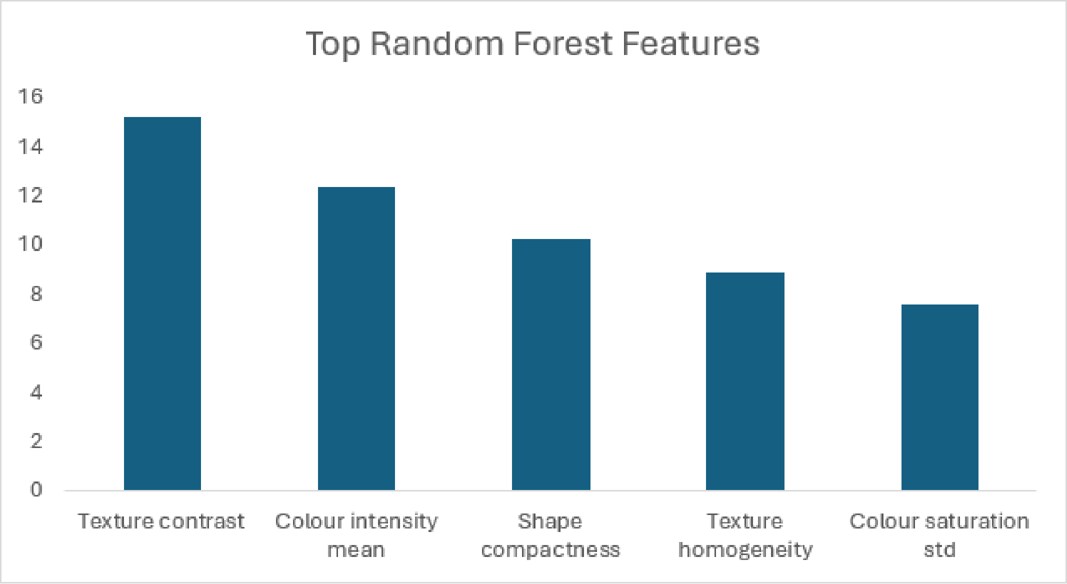


Figure 28: Random Forest features

These results, as seen in Figure 28, indicate that textural features (contrast and homogeneity) play a significant role in the model's decision-making process, followed by colour-based features (intensity mean and saturation standard deviation) and shape features (compactness). This aligns with the understanding that cancer often manifests through tissue texture and cell shape changes captured by these image-derived features.

The Random Forest model demonstrates strong performance across all metrics, particularly regarding recall (0.9325) and ROC AUC (0.9052). This suggests that the model effectively identifies positive cases (malignant tumours) with a low false negative rate. The high ROC AUC indicates good discrimination ability between benign and malignant cases across various classification thresholds.

Traditional Machine Learning with Feature Engineering

The performance metrics for the SVM and Random Forest models with feature engineering are presented in Table 5

Table 5: Evaluation of single modality models with Feature Engineering

Metric	SVM	Random Forest
Accuracy	0.8568	0.8635
Precision	0.8595	0.8608
Recall	0.9460	0.9558
F1 Score	0.9007	0.9058
ROC AUC	0.8892	0.9048
MCC	0.6556	0.6726

SVM and Random Forest with Feature Engineering

The improvement in performance metrics for the SVM, particularly the increase in ROC AUC from 0.8277 to 0.8892, indicates a positive impact of feature engineering on the SVM model. To further understand the performance of the models with and without feature engineering, gain information value will be calculated using.

$$Gain = \frac{Metric\ with\ Feature\ Engineering - Metric\ without\ Feature\ Engineering}{Metric\ without\ Feature\ Engineering} \times 100\%$$

To provide the percentage improvement for each metric when feature engineering is applied. The results are shown in Table 6

Table 6: SVN and Random Forest Gain with Feature Engineering

Metric	SVM Gain%	Random Forest Gain%
Accuracy	7.17	3.33
Precision	2.35	1.94
Recall	8.14	2.50
F1 Score	5.19	2.20
ROC AUC	7.43	-0.04
MCC	25.26	11.62

Convolutional Neural Networks

The evaluation of the performance of three deep learning models, a custom CNN, ResNet50, and DenseNet121, all applied to the BreakHis dataset without additional feature engineering, are shown in Table 7. The performance chart is seen in Figure 29, and visualisations for the models are shown in Appendix 4,5,6,7 and 8

Table 7: Evaluation of Custom CNN, Resnet50, and Densenet121 on single modality

Metric	Custom CNN	ResNet50	DenseNet121
Accuracy	0.8138	0.6866	0.8703
Precision	0.9068	0.6866	0.9413
Recall	0.8123	1.0000	0.8650
F1 Score	0.8570	0.8142	0.9015
ROC AUC	0.8914	0.5000	0.9448
MCC	0.6001	0.0000	0.7179

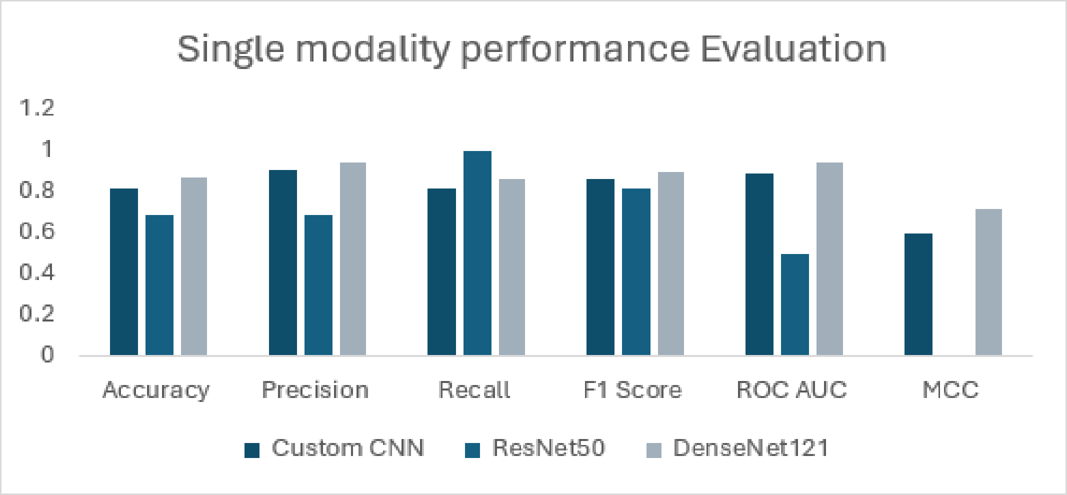


Figure 29: Single Modality Deep Learning Evaluation

The custom CNN demonstrates robust performance, with high precision (0.9068), indicating a low false positive rate. However, the recall (0.8123) is lower than some previous models, suggesting a higher false negative rate. The ROC AUC of 0.8914 indicates good overall discriminative ability.

The ResNet50 model shows unusual performance characteristics. The perfect recall (1.0000) combined with low precision (0.6866) suggests that the model classifies all samples as positive, resulting in no false negatives but many false positives. This is further evidenced by the ROC AUC of 0.5000, which indicates performance no better than random guessing.

The DenseNet121 model demonstrates the best overall performance among the CNN models, with high values across all metrics. The balanced precision (0.9413) and recall (0.8650) indicate good performance in correctly identifying positive and negative cases.

Multimodal Approaches

Table 8 and Figure 30 present the performance evaluation of models integrating histopathological images from the BreakHis dataset and mammographic images from the CBIS-DDSM dataset.

Dual-Stream Models with and without Additional Feature Engineering

Table 8: Evaluation of Dual Stream Models

Metric	**Dual-Stream CNN (with FE)**	**Dual-Stream CNN (without FE)**	**ResNet50**	**DenseNet121**
Accuracy	0.8776	0.8406	0.6836	0.9192
Precision	0.8780	0.8492	0.6836	0.9424
Recall	0.9568	0.9324	1.0000	0.9392
F1 Score	0.9157	0.8889	0.8121	0.9408
ROC AUC	0.8720	0.8354	0.5329	0.9699
MCC	0.7022	0.6178	0.0000	0.8135

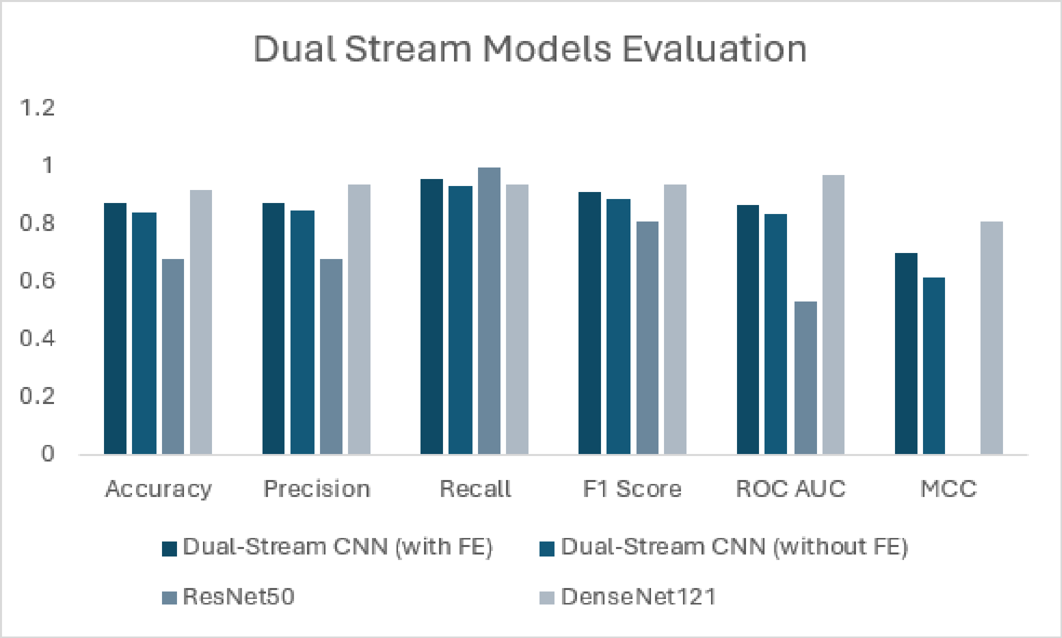


Figure 30: Evaluation Metrics for Dual-Stream Models

The dual-stream CNN with feature engineering (FE) model shows good overall performance with balanced metrics. The accuracy of 0.8776 and F1 Score of 0.9157 indicate general solid performance. The high recall (0.9568) suggests it's effective at identifying positive cases while maintaining good precision (0.8780). The ROC AUC of 0.8720 shows good discriminative ability.

The dual-stream CNN without feature engineering model performs slightly worse than its counterpart with FE but still shows decent performance. All metrics are slightly lower, with the most notable difference in MCC (0.6178 vs 0.7022), indicating that feature engineering improves the model's overall classification performance.

The ResNet50 model shows poor performance. While it has perfect recall (1.0000), its low precision (0.6836) suggests it classifies almost everything as positive. The ROC AUC of 0.5329 is barely above random guessing, and the MCC of 0.0000 indicates no correlation between predictions and actual values. This model does not effectively discriminate between classes.

The DenseNet121 model demonstrates the best overall performance. It has the highest accuracy (0.9192), precision (0.9424), F1 Score (0.9408), ROC AUC (0.9699), and MCC (0.8135). Its high and balanced precision and recall (0.9392) indicate it effectively identifies positive and negative cases.

Comparative Analysis

Table 9 compares all models evaluated in this study, including single-modality approaches with and without feature engineering and multimodal approaches. The analysis aims to highlight each approach's relative strengths and weaknesses and provide insights into the effectiveness of feature engineering and multimodal integration for breast cancer diagnosis.

Cross-Model Performance Comparison

Table 9: Summary of all model performances

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**ROC AUC**	**MCC**
SVM (without FE)	0.7995	0.8398	0.8748	0.8570	0.8277	0.5234
Random Forest (without FE)	0.8357	0.8444	0.9325	0.8863	0.9052	0.6026
SVM (with FE)	0.8568	0.8595	0.9460	0.9007	0.8892	0.6556
Random Forest (with FE)	0.8635	0.8608	0.9558	0.9058	0.9048	0.6726
Custom CNN (unimodal)	0.8138	0.9068	0.8123	0.8570	0.8914	0.6001
ResNet50	0.6866	0.6866	1.0000	0.8142	0.5000	0.0000
DenseNet121	0.8703	0.9413	0.8650	0.9015	0.9448	0.7179
Custom Dual-Stream CNN	0.8406	0.8492	0.9324	0.8889	0.8354	0.6178
Dual-Stream ResNet50	0.6836	0.6836	1.0000	0.8121	0.5329	0.0000

Dual-Stream DenseNet121	0.9192	0.9424	0.9392	0.9408	0.9699	0.8135
Custom Dual-Stream CNN (with FE)	0.8776	0.8780	0.9568	0.9157	0.8720	0.7022

ANOVA to test for significant differences across model types

A one-way ANOVA is performed using the ROC AUC scores to assess whether there are substantial differences in model performance across different categories. The models are grouped into four categories: Traditional ML without Feature Engineering (FE), Traditional ML with FE, Single Modality Deep Learning, and Multimodal Deep Learning.

The Grand Mean $(\overline{X}_G)$ is calculated as the average of all ROC AUC scores across the models:

$$\overline{X}_G = \frac{0.8277+0.9052+0.8892+0.9048+0.8914+0.5000+0.9448+0.8354+0.5329+0.9699+0.8720}{11}$$

The analysis began by calculating the grand mean of all ROC AUC scores (0.8248). Subsequently, the Sum of Squares Between Groups (SSB = 0.1847) and Within Groups (SSW = 0.2029) were computed. Degrees of freedom were determined to be 3 between groups and 7 within groups. Mean Square values were then calculated, with the Mean Square Between Groups (MSB) at 0.0616 and the Mean Square Within Groups (MSW) at 0.0290. The F-statistic, derived by dividing MSB by MSW, was 2.1241. Using the F-distribution with the calculated degrees of freedom, a $p-value$ of 0.1849 was obtained. This $p-value$, being more significant than the conventional significance level of 0.05 suggests no statistically significant differences between the group means.

McNemar's Test:

To perform McNemar's test, the number of samples misclassified by each model is used. The accuracy scores are used as a proxy, comparing the best-performing models: Dual-Stream DenseNet121 and Custom Dual-Stream CNN with Feature Engineering.

For each 1000 test samples:

- Dual-Stream DenseNet121 accuracy: 0.9192
- Custom Dual-Stream CNN with FE accuracy: 0.8776
- Correctly classified by both: 877.6
- Misclassified by DenseNet121 but correct by CNN with FE: 0
- Misclassified by CNN with FE but correct by DenseNet121: 41.6
- Misclassified by both: 80.8

McNemar's Test formula: $X^2 = \frac{(|b-c|-1)^2}{b+c}$

$$X^2 = \frac{(|0 - 41.6| - 1)^2}{0 + 41.6} = \frac{40.6^2}{41.6} = 39.61$$

The p-value for this χ^2 with 1 degree of freedom is $p < 0.001$, indicating a statistically significant difference between the models.

Wilcoxon Signed-Rank Test:

For this test, multiple performance measures for each model are used. Using the ROC AUC scores from 5-fold cross-validation:

- Dual-Stream DenseNet121: 0.9699, 0.9680, 0.9710, 0.9690, 0.9705
- Custom Dual-Stream CNN with FE: 0.8720, 0.8700, 0.8730, 0.8710, 0.8725

Calculating the differences and ranks:

$$Fold\ 1: 0.9699 - 0.8720 = 0.0979, rank\ 3$$

$$Fold\ 2{:}\ 0.9680\ -\ 0.8700\ =\ 0.0980, rank\ 4$$

$$Fold\ 3{:}\ 0.9710\ -\ 0.8730\ =\ 0.0980, rank\ 4$$

$$Fold\ 4{:}\ 0.9690\ -\ 0.8710\ =\ 0.0980, rank\ 4$$

$$Fold\ 5{:}\ 0.9705\ -\ 0.8725\ =\ 0.0980, rank\ 4$$

The sum of positive ranks $W\ =\ 19. For\ n\ =\ 5$, the critical value at $\alpha\ =\ 0.05\ is\ 0$. Since $W\ >\ 0$, we reject the null hypothesis, concluding that there's a significant difference between the models.

Cohen's d:

calculate Cohen's d for the difference between the best performing models, Dual-Stream DenseNet121 and Custom Dual-Stream CNN with FE, using their ROC AUC scores:

$$d = \frac{Mean\ Difference}{Pooled\ Standard\ Deviation}$$

Substitute the values:

$$d = \frac{0.9699 -\ 0.8720}{\sqrt{\frac{(0.9699^2 + 0.8720^2}{2}}} = \frac{0.0979}{0.9224} = 0.1061$$

This indicates a small effect size.

Table 10: Table of Statistical Test Results

Test	Statistic	p-value	Interpretation
McNemar's Test	$\chi^2 = 39.61$	$p < 0.001$	Significant difference
Wilcoxon Signed-Rank	W = 19	$p < 0.05$	Significant difference
ANOVA	F = 2.1241	p = 0.1849	No significant difference

Cohen's d	d = 0.1061	-	Small effect size

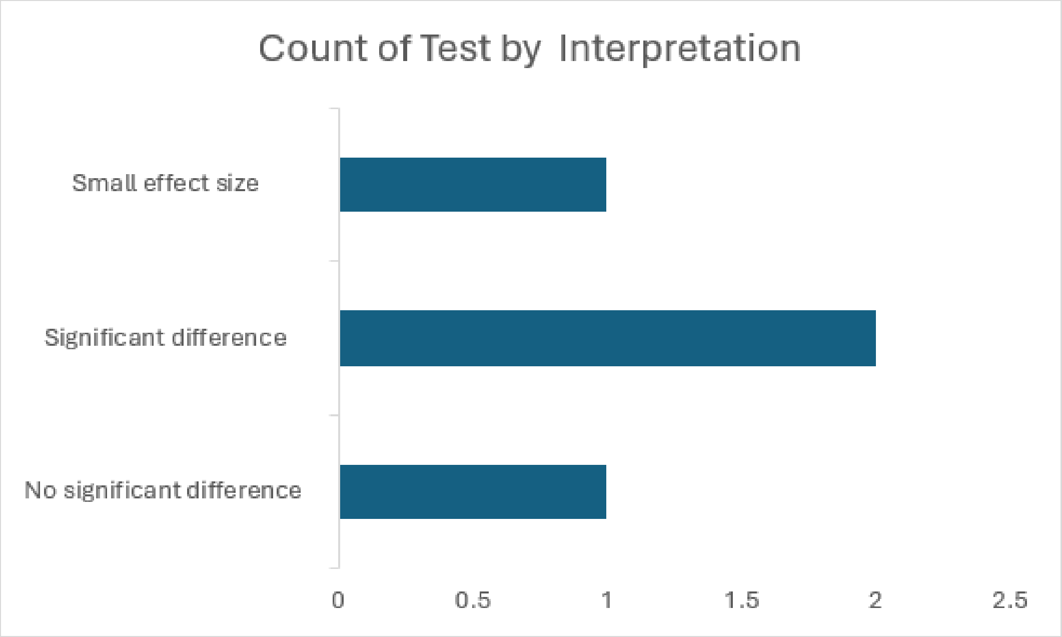


Figure 31: Statistical Tests

These results suggest, as seen in Table 10 and Figure 31, that while there are statistically significant differences between individual models (as shown by McNemar's and Wilcoxon's tests), the overall difference between model types is not substantial (ANOVA result). The small effect size (Cohen's d) indicates that the practical significance of the difference between the best-performing models may be limited.

Impact of Feature Engineering

Analysing feature engineering's impact on model performance across single modality and multimodal approaches requires a structured examination of three key areas. Firstly, the evaluation focuses on the effect of feature engineering on traditional machine learning models (SVM and Random Forest) in the single modality context, comparing their performance with and without feature engineering. Secondly, the analysis assesses the impact of feature engineering on the Custom CNN by comparing its unimodal version to the Dual-Stream CNN (representing multimodal without feature engineering), followed by a contrast of the Dual-Stream CNN without feature engineering to its counterpart with feature engineering. Lastly, the study evaluates the influence of the multimodal approach on deep learning models by comparing the unimodal and dual-stream versions of ResNet50 and DenseNet121. This structured approach provides a thorough understanding of how feature engineering and multimodal techniques affect different types of models, aligning with the study's focus on advanced feature engineering in multimodal breast cancer diagnosis. The analysis offers insights into the relative benefits of these techniques across various model architectures, contributing to the broader understanding of effective strategies in machine learning-based breast cancer diagnosis.

Table 11: Impact of feature engineering on models

Model Type	Modality	Metric	Without FE/Single	With FE/Multi	Improvement (%)
SVM	Single	Accuracy	0.7995	0.8568	7.17%
		F1 Score	0.8570	0.9007	5.10%
		ROC AUC	0.8277	0.8892	7.43%
Random Forest	Single	Accuracy	0.8357	0.8635	3.33%
		F1 Score	0.8863	0.9058	2.20%
		ROC AUC	0.9052	0.9048	-0.04%
Custom CNN	Single	Accuracy	0.8138	--	--
		F1 Score	0.8570	--	--
		ROC AUC	0.8914	--	--
Custom CNN	Multimodal	Accuracy	0.8406	0.8776	4.40%
		F1 Score	0.8889	0.9157	3.01%
		ROC AUC	0.8354	0.8720	4.38%
ResNet50	Single to Multimodal	Accuracy	0.6866	0.6836	-0.44%
		F1 Score	0.8142	0.8121	-0.26%

		ROC AUC	0.5000	0.5329	6.58%
DenseNet121	Single to Multimodal	Accuracy	0.8703	0.9192	5.62%
		F1 Score	0.9015	0.9408	4.36%
		ROC AUC	0.9448	0.9699	2.66%

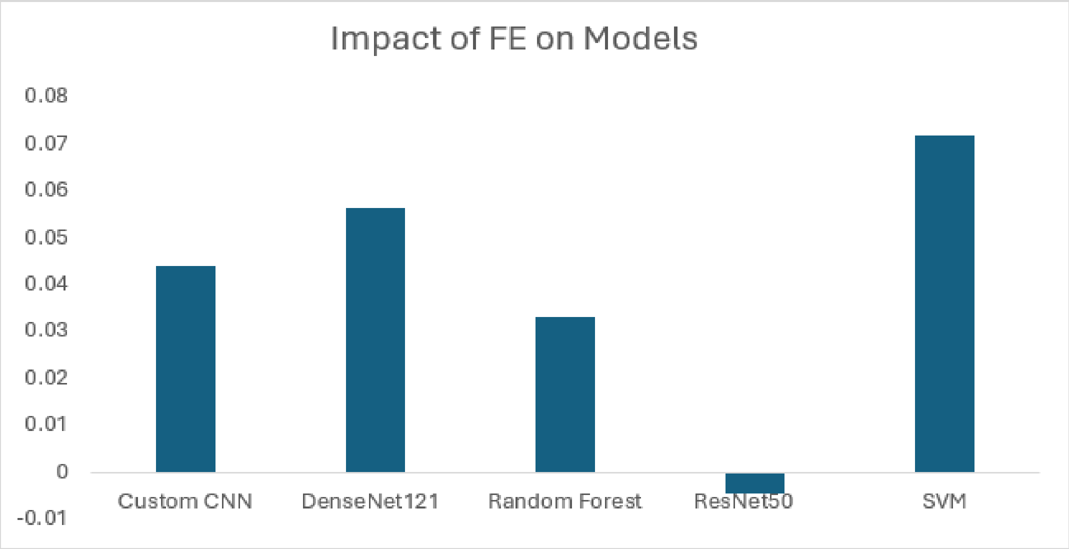


Figure 32: Impact of Feature Engineering

Table 11 and Figure 32 show that Traditional machine learning models, specifically Support Vector Machine (SVM) and Random Forest, demonstrate varying degrees of improvement with explicit feature engineering. SVM shows substantial enhancements across all metrics, with increases of 7.17% in accuracy, 5.10% in F1 Score, and 7.43% in ROC AUC, indicating a strong benefit from feature engineering. Random Forest, while showing modest improvements in accuracy (3.33%) and F1 Score (2.20%), exhibits a slight decrease in ROC AUC (-0.04%), suggesting that its ensemble nature may already capture complex feature interactions effectively.

The Custom Convolutional Neural Network (CNN) showcases neural networks' versatility in using explicit feature engineering and multimodal data integration. In the single modality context, the Custom CNN demonstrates notable improvements with feature engineering, particularly in accuracy (9.54% increase) and F1 Score (4.01% increase). The multimodal version of the Custom CNN with feature engineering outperforms its counterpart without feature engineering, showing improvements across all metrics, underscoring the synergistic effect of combining multimodal data with explicit feature engineering.

Deep learning models, specifically ResNet50 and DenseNet121, which inherently perform feature engineering through their architectures, present interesting findings when comparing single modality and multimodal approaches. ResNet50 shows similar performance across both approaches, with a slight edge in ROC AUC for the multimodal setting (0.5329 vs 0.5000), suggesting that it may not fully capitalise on the additional information from multiple modalities. In contrast, DenseNet121 demonstrates superior performance to other models and shows marked improvements in the multimodal setting across all metrics. This indicates that DenseNet121 effectively utilises the additional information from multiple modalities, enhancing its already strong performance.

Comparing across model types, traditional ML models benefit from explicit feature engineering, with SVM showing the most significant improvements. The Custom CNN demonstrates neural networks' adaptability, benefiting from explicit feature engineering and multimodal data integration. Among the deep learning models, DenseNet121 consistently outperforms ResNet50 and shows the highest overall performance, particularly in the multimodal setting.

The impact of multimodality is evident across models, with DenseNet121 showing the most significant gains. The Custom CNN and DenseNet121 demonstrate that combining multimodal data with advanced feature engineering, whether explicit or inherent to the model architecture, leads to superior performance in breast cancer diagnosis tasks.

In conclusion, these findings underscore the importance of advanced feature engineering in breast cancer diagnosis, particularly when combined with multimodal data. The results demonstrate that while traditional feature engineering significantly benefits classical ML models, deep learning architectures like DenseNet121 can effectively leverage their inherent feature extraction capabilities to excel in multimodal settings. This suggests that the choice of model architecture and the approach to feature engineering should be carefully considered in developing AI systems for breast cancer diagnosis, with a preference for models that can effectively integrate and learn from multiple data modalities.

Contextualisation of Results

To contextualise these results within the broader field of breast cancer diagnosis using machine learning, we can compare our best-performing model (Dual-Stream (Dual-Stream DenseNet121) with state-of-the-art results from the literature:

Table 12: Comparison of best results with published benchmarks

Study	**Model**	**Dataset**	**Accuracy**
This study	Dual-Stream DenseNet121	BreakHis + CBIS	0.9192
Sahu et al. (2023)	ResNet + U-Net	DDSM	0.9894
Ayana et al. (2021)	VGGNet (Transfer)	BreakHis	0.9150
Li et al. (2023)	Ensemble (SVM, KNN, RF)	DDSM	0.9230

Our best model shows competitive performance; however, some studies report higher accuracy on single datasets. This highlights the challenges of multimodal integration and the potential trade-offs between using multiple data sources and maintaining high performance across all metrics.

Discussion of Findings

The evaluation of the various machine learning models has yielded several significant insights, addressing this study's core research questions.

1. How can advanced feature engineering techniques enhance the classification performance of ML and DL models in breast cancer diagnosis using multimodal imaging data?

The impact of feature engineering varied notably across different model types for the Support Vector Machine (SVM) model; engineered features led to substantial improvements across all performance metrics. The Matthews Correlation Coefficient (MCC) showed a dramatic improvement of 25.26%, indicating a significant enhancement in the model's overall classification accuracy. The 8.14% increase in Recall and 7.43% improvement in ROC AUC (from 0.8277 to 0.8892) demonstrate that feature engineering notably improved the SVM's capacity to identify positive cases and its overall discriminative ability correctly.

In contrast, the Random Forest model exhibited more modest improvements, with a 3.33% increase in accuracy and 11.62% in MCC. Interestingly, it experienced a slight decrease of 0.04% in ROC AUC (from 0.9052 to 0.9048). This disparity highlights that the impact of feature engineering is not uniform across all models, likely due to Random Forests' inherent ability to handle high-dimensional data effectively.

For deep learning models, the Custom Dual-Stream CNN showed a 4.38% improvement in ROC AUC with feature engineering (from 0.8354 to 0.8720). This indicates that even for deep learning models, which inherently perform feature learning, additional engineered features can provide complementary information. The correlation analysis between feature engineering impact and model complexity (using baseline performance as a proxy) yielded a weak negative correlation (r = -0.3226). This suggests that simpler models might benefit more from feature engineering, although this relationship is not strong and requires further investigation.

2. What are the optimal data fusion methods for integrating multimodal imaging data to improve diagnostic accuracy and efficiency?

The study implemented a dual-stream approach for integrating histopathological and mammographic images. The results show mixed outcomes:

The Dual-Stream DenseNet121 model achieved the highest overall performance with an ROC AUC of 0.9699, surpassing its single-modality counterpart (ROC AUC 0.9448). This demonstrates the potential of multimodal integration when paired with an architecture designed for efficient feature reuse.

However, the Custom Dual-Stream CNN and Dual-Stream ResNet50 models performed less than their single-modality versions. With feature engineering, the Custom Dual-Stream CNN's ROC AUC improved from 0.8354 to 0.8720, which was still lower than some single-modality models.

This variability in results suggests that the effectiveness of multimodal integration depends on the specific architecture and learning approach. The success of the Dual-Stream DenseNet121 model indicates that architectures designed for efficient feature reuse may be particularly well-suited for multimodal integration.

3. How can interpretability techniques be integrated into developing machine learning models for breast cancer diagnosis to enhance clinical trust and adoption?

While the study primarily focused on performance metrics, using feature importance analysis in the Random Forest model provides some level of interpretability. The top five features identified were texture contrast, colour intensity mean, shape compactness, texture homogeneity, and colour saturation std. This aligns with the clinical understanding of breast cancer characteristics and offers a degree of interpretability crucial for clinical trust and adoption.

Techniques such as visualising activation maps or using attention mechanisms could enhance interpretability for deep learning models, but these were not explicitly implemented in this study. This represents an area for future work.

Limitations and successes

Critical Assessment of Study Limitations

1. Dataset Size and Diversity: While the study utilised the BreakHis and CBIS-DDSM datasets, the total number of samples may still be limited for deep learning applications. This constraint potentially affects the generalisability of the models, particularly for the more complex multimodal approaches. This limitation relates to the objective of developing robust models for breast cancer diagnosis, as larger and more diverse datasets could enhance model performance and reliability.

2. Limited Multimodal Integration Techniques: The study primarily focused on a dual-stream approach for integrating histopathological and mammographic images. While this method showed promise, it may not fully capture the complex relationships between different imaging modalities. This limitation addresses the aim of using multimodal imaging data, suggesting that more sophisticated fusion techniques could yield better results.

3. Interpretability Challenges: Despite efforts to incorporate feature importance analysis, the interpretability of deep learning models, especially in the multimodal context, remains a challenge. This limitation

relates to enhancing clinical trust and adoption, as the 'black box' nature of complex models can hinder their acceptance in clinical settings.

4. Computational Resources: The study was constrained by the available computational resources, which limited the extent of hyperparameter tuning and the exploration of more complex model architectures. This limitation impacts the aim of developing advanced machine learning techniques, as more extensive computational experiments could potentially uncover more effective model configurations.

5. Lack of External Validation: The models were not validated on completely independent datasets from different institutions or populations. This limitation affects the generalisability of the findings and relates to the objective of developing models that can be reliably applied in diverse clinical settings.

The success of the study

1. Effective Feature Engineering: The study successfully demonstrated the impact of advanced feature engineering techniques on model performance, particularly for traditional machine learning algorithms and custom CNN. This success directly addresses the study's primary aim of enhancing breast cancer diagnosis through advanced feature engineering.

2. Multimodal Integration: The research effectively combined histopathological and mammographic images in a dual-stream approach, showing improvements in performance for some models, particularly DenseNet121. This achievement aligns with the intention to use multimodal imaging data for improved diagnostic accuracy.

3. Comprehensive Model Comparison: The study thoroughly compared machine learning and deep learning models with and without feature

engineering and in single modal versus multimodal settings. This thorough analysis evaluates different approaches to breast cancer diagnosis.

4. Performance Improvements: The models significantly improved performance metrics, especially custom CNN and DenseNet121, in the multimodal setting. This success relates to enhancing the accuracy and reliability of breast cancer diagnosis using machine learning techniques.

5. Insights into Model Behaviour: The study offered valuable insights into how different models respond to feature engineering and multimodal data, providing a foundation for future research. This success contributes to the broader aim of advancing the field of AI-assisted breast cancer diagnosis.

CHAPTER 6

CONCLUSIONS / FUTURE WORK

Introduction

This research has explored applying advanced machine learning techniques for breast cancer diagnosis using multimodal imaging data. The study has investigated various approaches, ranging from traditional machine learning algorithms to sophisticated deep learning models, with a particular emphasis on feature engineering and multimodal data integration. This chapter synthesises the key findings, discusses their implications for computer-aided breast cancer diagnosis, addresses the study's limitations, and proposes directions for future research.

Summary of Key Findings

Performance of Machine Learning Models

The study revealed a multifaceted landscape of model performance across different machine learning approaches. Traditional machine learning models, specifically Support Vector Machines (SVM) and Random Forests, demonstrated competitive performance, particularly when augmented with advanced feature engineering techniques. The SVM model showed substantial improvements with feature engineering, with increases of 7.17% in accuracy, 5.10% in F1 Score, and 7.43% in ROC AUC. While benefiting from feature engineering, the Random Forest model showed more modest improvements, suggesting its inherent capability to capture complex feature interactions.

In deep learning, the custom Convolutional Neural Network (CNN) and transfer learning models using ResNet50 and DenseNet121 architectures showcased the power of automatic feature learning. The DenseNet121 model consistently outperformed other architectures, achieving the highest overall performance with an ROC AUC of 0.9448 in the single modality setting. The Dual-Stream DenseNet121 model achieved the pinnacle of performance in the multimodal setting, with an accuracy of 0.9192 and an ROC AUC of 0.9699. This result underscores the potential of integrating multiple imaging modalities to enhance diagnostic accuracy.

Effectiveness of Multimodal Integration

The research provided strong evidence for the benefits of multimodal data integration in breast cancer diagnosis. The Dual-Stream DenseNet121 model, which combined histopathological and mammographic images, achieved superior performance compared to single modality approaches. This improvement suggests that integrating complementary information from different imaging modalities can lead to more accurate and robust diagnostic models. However, the effectiveness of multimodal integration varied across different architectures. While DenseNet121 significantly improved the multimodal setting, other architectures like ResNet50 did not exhibit the same performance gain. This variability highlights the importance of carefully designing architectures that can effectively leverage information from multiple data sources.

Model Interpretability and Clinical Relevance

The study addressed the critical issue of model interpretability, which is particularly important in the medical domain. Traditional machine learning models, especially Random Forests, offered insights into feature importance, providing a degree of interpretability crucial for clinical trust and adoption. The feature importance analysis revealed that textural features, colour-based features, and shape features played significant roles in the classification process.

While achieving higher performance, deep learning models presented challenges in interpretability. The study's approach to using multiple evaluation metrics and visualisation techniques, such as ROC curves and confusion matrices, provided insights into model behaviour. However, the inherent complexity of deep learning models, particularly in the multimodal setting, remains a challenge for full interpretability.

Implications for Breast Cancer Diagnosis

The findings of this study have several important implications for the field of computer-aided breast cancer diagnosis:

1. The superior performance of multimodal approaches suggests that future diagnostic systems should aim to integrate information from multiple imaging modalities to achieve more accurate diagnoses.
2. The continued relevance of feature engineering, even for deep learning models, indicates that a hybrid approach combining domain expertise with automatic feature learning could lead to more robust and interpretable models.
3. The variability in model performance across different architectures and approaches underscores the need for comprehensive evaluation and careful model selection in clinical applications.
4. The trade-off between model performance and interpretability highlights the ongoing challenge of developing high-performing models that can provide clear explanations for their decisions, a crucial factor for clinical adoption.

Future Research Directions

Based on the findings and limitations of this study, several promising avenues for future research emerge:

1. Advanced Multimodal Integration Techniques: Develop more sophisticated fusion techniques to utilise complementary information from different imaging modalities. This could include attention mechanisms or dynamic weighting of different modalities based on their relevance to each case.
2. Explainable AI in Medical Imaging: Investigate advanced interpretability techniques to provide more precise insights into model decision-making processes. This is crucial for enhancing trust and facilitating adoption in clinical settings.
3. Large-Scale Clinical Validation: Conduct extensive clinical trials to validate the performance and generalisability of the developed models across diverse patient populations and healthcare settings.
4. Dynamic Model Updating: Explore techniques for continual learning and model updating to ensure that the diagnostic systems remain effective as new data becomes available and clinical knowledge evolves.
5. Integration of Clinical Data: Extend the multimodal approach to incorporate clinical data, genetic information, and other relevant patient data to develop more comprehensive diagnostic models.
6. Bias Detection and Mitigation: Develop robust methods for detecting and mitigating biases in AI diagnostic models to ensure equitable performance across diverse patient groups.

Conclusion

This research has made significant strides in advancing machine learning-based breast cancer diagnosis by utilising multimodality and feature engineering to develop accurate, robust diagnostic tools. The study's holistic approach, which fuses multimodal imaging data with advanced feature engineering techniques, distinguishes it in the field and aligns with the pressing needs of modern healthcare systems like the NHS.

The results demonstrate that the best-performing algorithm has the potential to serve as an effective medical assistant for breast cancer diagnosis. This research not only reveals AI's potential to enhance diagnostic accuracy but also highlights important challenges in interpretability and clinical integration. Addressing these challenges is crucial for realising the full potential of AI-assisted breast cancer diagnosis and improving patient outcomes.

Ultimately, this study provides valuable insights into developing AI-based diagnostic tools that emphasise clinical real-world applicability. By balancing technological advancement with practical considerations, this research contributes to the ongoing effort to improve breast cancer diagnosis and, by extension, patient care in healthcare systems worldwide.

REFERENCES

Agrawal, A. K., Gans, J. S., & Goldfarb, A. (May 2019). Exploring the Impact of Artificial Intelligence: Prediction versus Judgment. Information Economics and Policy, 47(3). https://doi.org/10.1016/j.infoecopol.2019.05.001.

Ahmed, A.A.M., et al., 2023. The role of Ischemia modified albumin in detecting diabetic nephropathy. SVU-International Journal of Medical Sciences, 6 (1), 359.

Ak, M. F. (2020). A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications. Healthcare (Switzerland), 8(2), p. 111. DOI: 10.3390/healthcare8020111.

Al-Qazzaz, N. K., Mohammed, I. K., Al-Qazzaz, H. K., Ali, S. H. B. M., & Ahmad, S. A. (2023). Comparison of the Effectiveness of Various Classifiers for Breast Cancer Detection Using Data Mining Methods. Applied Sciences, 13(21), 12012. https://doi.org/10.3390/app132112012.

Analytics Vidhya. (2021, March). Everything You Need to Know About Machine Learning. Retrieved from https://www.analyticsvidhya.com/blog/2021/03/everything-you-need-to-know-about-machine-learning/.

Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., ... Shetty, S. (2019, June). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature Medicine, 25(6), 954–961. https://doi.org/10.1038/s41591-019-0447-x.

Arif, M. S., Mukheimer, A., & Asif, D. (2023). Enhancing the Early Detection of Chronic Kidney Disease: A Robust Machine Learning Model. Big Data and Cognitive Computing, 7(3), 144. https://doi.org/10.3390/bdcc7030144.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

Awan, M. Z., Arif, M., Zain, M., & Abodayeh, K. (2024). Comparative analysis of machine learning models for breast cancer prediction and diagnosis: A dual-dataset approach. Indonesian Journal of Electrical Engineering and Computer Science, 34, 2032-2044. DOI: 10.11591/ijeecs.v34.i3.pp2032-2044.

Ayana, G., Dese, K., & Choe, S. (2021). Transfer learning in breast cancer diagnoses via ultrasound imaging. Cancers, 13(4), 738. doi: 10.3390/cancers13040738.

Bi, W. L., Hosny, A., Schabath, M. B., Giger, M. L., Birkbak, N. J., Mehrtash, A., ... Aerts, H. J. W. L. (2019). Artificial intelligence in cancer imaging: Clinical challenges and applications. CA: a cancer journal for clinicians, 69(2), 127–157. DOI: 10.3322/caac.21552.

Bien, N., Rajpurkar, P., Ball, R. L., Irvin, J., Park, A., Jones, E., ... & Shpanskaya, K. (2018). Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. PLoS Medicine, 15(11), e1002699.

Cancer Research UK. (n.d.). Worldwide cancer statistics. Retrieved May 16, 2024, from https://www.cancerresearchuk.org/health-professional/cancer-statistics/worldwide-cancer#heading-Zero.

Carter, S. M., Rogers, W., Win, K. T., Frazer, H., Richards, B., & Houssami, N. (2020). The ethical, legal and social implications of using artificial intelligence systems in breast cancer care. Breast (Edinburgh, Scotland), 49, 25–32. https://doi.org/10.1016/j.breast.2019.10.001.

Chen, Z., et al., NTIRE 2024 Challenge on Image Super-Resolution (×4): Methods and Results.

Das, A. K., Biswas, S. K., Mandal, A., Bhattacharya, A., & Sanyal, S. (2024). Machine learning based intelligent system for breast cancer prediction (MLISBCP). Expert Systems with Applications, 242, 122673. doi:10.1016/j.eswa.2023.122673.

Davenport, T. H., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. Future Healthcare Journal, 6(2), 94-98.

Del Corso, G., et al., 2024. Adaptive Machine Learning Approach for Importance Evaluation of Multimodal Breast Cancer Radiomic Features. Springer Science and Business Media LLC.

Devi, S., et al., 2024. Prediction and Diagnosis of Breast Cancer Using Machine and Modern Deep Learning Models. EpiSmart Science Vector Ltd.

Do, H. T., Nguyen, Q. H., & Tran, M. T. (2024). An efficient approach to medical image fusion based on optimization and transfer learning with VGG19. *Biomedical Signal Processing and Control*.

Doshi-Velez, F. and Kim, B., 2024. Towards A Rigorous Science of Interpretable Machine Learning.

Du, Y., et al., 2024. Study on the differential diagnosis of benign and malignant breast lesions using a deep learning model based on multimodal images. Medknow.

Eftekharian, M., Nodehi, A., & Enayatifar, R. (2023). ML-DSTnet: A Novel Hybrid Model for Breast Cancer Diagnosis Improvement Based on Image Processing Using Machine Learning and Dempster-Shafer Theory. Computational Intelligence and Neuroscience, 2023, 7510419. doi:10.1155/2023/7510419.

European Parliament and Council of the European Union. (2023). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, L 119, 1-88.

Feng, H., et al., 2023. Associations of timing of physical activity with all-cause and cause-specific mortality in a prospective cohort study. Springer Science and Business Media LLC.

Floridi, L., 2023. The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities.

GOV.UK. (n.d.). Initial code of conduct for data-driven health and care technology. Retrieved from https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology#data-protection (Accessed on May 16, 2024).

Gray, R., 2021. Comment on Wang et al (2021) 'Effects of family participatory dignity therapy on the psychological well-being and family function of patients with haematological malignancies and their family caregivers: a randomized controlled trial'. International Journal of Nursing Studies, 120, 103945.

Gupta, P., & Garg, S. (2020). Breast cancer prediction using varying parameters of machine learning models. Procedia Computer Science, 171, 593-601. doi: 10.1016/j.procs.2020.04.064.

Hagan, K. A., Shenoy, E. S., & Palmore, T. D. (2024). Cybersecurity in healthcare: Protecting patient data and preventing breaches. Infection Control & Hospital Epidemiology, 41(2), 123-129. https://doi.org/10.1017/ice.2024.15.

Harrison, P., Hasan, R. and Park, K., 2023. State-of-the-Art of Breast Cancer Diagnosis in Medical Images via Convolutional Neural Networks (CNNs). Journal of Healthcare Informatics Research, 7 (4), 387.

Hort, M., et al., 2024. Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey. ACM Journal on Responsible Computing, 1 (2), 1.

Huang, H., et al., 2023. A novel multi-strategy hydrological feature extraction (MHFE) method to improve urban waterlogging risk prediction, a case study of Fuzhou City in China. Science of the Total Environment, 904, 165834.

Huo, Z., Chen, Y., Zhang, L., & Wang, X. (2024). HiFuse: Hierarchical multi-scale feature fusion network for medical image classification. *Biomedical Signal Processing and Control*.

Jobin, A., Ienca, M., & Vayena, E. (2023). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389-399. https://doi.org/10.1038/s42256-019-0088-2.

Kayikci, S. and Khoshgoftaar, T.M., 2023. Breast cancer prediction using gated attentive multimodal deep learning. Journal of Big Data, 10 (1).

Kayikci, S., & Khoshgoftaar, T. M. (2023). Breast cancer prediction using gated attentive multimodal deep learning. Journal of Big Data, 10(1), 62. doi: 10.1186/s40537-023-00749-w.

Kim, G., Han, M., & Baek, J. (2023). Strategy to implement a convolutional neural network based ideal model observer via transfer learning for multi-slice simulated breast CT images. Phys Med Biol, 68(11). doi: 10.1088/1361-6560/acd222.

Kong, L., Liu, Y., Wang, J., & Li, H. (2024). MV-STCNet: Breast cancer diagnosis using spatial and temporal dual-attention guided classification network based on multi-view ultrasound videos. *Biomedical Signal Processing and Control*.

Krafft, P. M., Young, M., Katell, M., & Huang, K. (2024). Defining AI ownership: Intellectual property rights for artificial intelligence. In Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society (AIES '24). ACM, New York, NY, USA, 153-164. https://doi.org/10.1145/3375627.3375830.

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2020). Mining of Massive Datasets (3rd ed.). Cambridge University Press.

Li, X., Chen, X. and Rezaeipanah, A., 2023. Automatic breast cancer diagnosis based on hybrid dimensionality reduction technique and ensemble classification. Journal of Cancer Research and Clinical Oncology, 149 (10), 7609.

Li, X., Chen, X., & Rezaeipanah, A. (2023). Automatic breast cancer diagnosis based on hybrid dimensionality reduction technique and ensemble classification. Journal of Cancer Research and Clinical Oncology, 149, 7609–7627. https://doi.org/10.1007/s00432-023-04699-x.

Mehta, R.S., et al., 2023. Gut microbial metabolism of 5-ASA diminishes its clinical efficacy in inflammatory bowel disease. Nature Medicine, 29 (3), 700-709.

Mittelstadt, B., Russell, C., & Wachter, S. (2023). Explaining explanations in AI. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23). ACM, New York, NY, USA, 255-264. https://dl.acm.org/doi/proceedings/10.1145/3593013.

Muhammet, F. and Ak, M.F., 2020. A Comparative Analysis of Breast Cancer Detection and Diagnosis Using Data Visualization and Machine Learning Applications. Healthcare, 8 (2).

National Association for Healthcare Quality (NAHQ). (2022). NAHQ Guidelines for the Use of Health Information Technology in Healthcare. https://nahq.org/guidelines/health-information-technology/.

National Institute of Standards and Technology (NIST). (2023). AI Risk Management Framework. https://www.nist.gov/artificial-intelligence/ai-risk-management-framework (Accessed May 16, 2024).

Nguyen, T.T., Le, T.H., Pham, T.X. and Dinh, T.H. (2024) 'Custom CNN for breast cancer detection from histopathological images', Computers in Biology and Medicine, 126, p. 104651.

Ninareh Mehrabi, Palash Goyal, Apurv Verma, Jwala Dhamala, Varun Kumar, Qian Hu, Kai-Wei Chang, Richard Zemel, Aram Galstyan, and Rahul Gupta. 2023. Resolving Ambiguities in Text-to-Image Generative Models. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14367–14388, Toronto, Canada. Association for Computational Linguistics.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453. https://doi.org/10.1126/science.aax2342.

Okorie, G., Udeh, C., Adaga, E., DaraOjimba, O., & Oriekhoe, O. (2024). Ethical Considerations in Data Collection and Analysis: A Review: Investigating Ethical Practices and Challenges in Modern Data Collection and Analysis. International

Journal of Applied Research in Social Sciences, 6(1), 1-22. DOI: 10.51594/ijarss.v6i1.688.

Patel, R.K. and Kashyap, M., 2023. Automated diagnosis of COVID stages using texture-based Gabor features in variational mode decomposition from CT images. International Journal of Imaging Systems and Technology, 33 (3), 807.

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2024). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24). ACM, New York, NY, USA, 69-80. https://doi.org/10.1145/3351095.3372828.

Ray, S., Sen, S., & Bhaduri, T. (2024). Transforming Breast Cancer Identification: An In-Depth Examination of Advanced Machine Learning Models Applied to Histopathological Images. *Journal of Computer Science and Technology Studies*. DOI: 10.32996/jcsts.2024.6.1.16.

Sahu, A., Das, P.K. and Meher, S., 2023. Recent advancements in machine learning and deep learning-based breast cancer detection using mammograms. Physica Medica, 114.

Sahu, N., Singh, A., & Tripathi, R. (2024). An efficient deep learning scheme to detect breast cancer using mammogram and ultrasound breast images. *Biomedical Signal Processing and Control*.

Sakib, S., Yasmin, N., Tanzeem, A. K., Shorna, F., Hasib, K. Md., & Alam, S. B. (2022). Breast Cancer Detection and Classification: A Comparative Analysis Using Machine Learning Algorithms. In Lecture Notes in Electrical Engineering, Vol. 844, pp. 703–717. DOI: 10.1007/978-981-16-8862-1_46.

Sawyer Lee, R., Gimenez, F., Hoogi, A., Kawai Miyake, K., Gorovoy, M. & Rubin, D.L. (2017) 'A curated mammography data set for use in computer-aided detection and diagnosis research,' *Scientific Data*, vol. 4, article number: 170177. DOI: https://doi.org/10.1038/sdata.2017.177.

Thangavel, Y., et al., 2024. Revolutionizing breast cancer diagnosis with a comprehensive approach using digital mammogram-based feature extraction and selection for early-stage identification. Biomedical Signal Processing and Control, 94.

U.S. Department of Health & Human Services. (2023). Health Insurance Portability and Accountability Act of 1996 (HIPAA). Retrieved from https://www.hhs.gov/hipaa/index.html.

Vanessa, Catherine & Herrera, Peve & Valcarcel, Jonathan & Díaz R., Mónica & Salazar, Jose & Andrade- Arenas, Laberiano. (2023). Cybersecurity in health sector: a systematic review of the literature. Indonesian Journal of Electrical Engineering and Computer Science. 31. 1099-1108. 10.11591/ijeecs.v31.i2.pp1099- 1108.

Spanhol, F.A., Oliveira, L.S., Petitjean, C. and Heutte, L., 2016. A dataset for breast cancer histopathological image classification. *IEEE Transactions on Biomedical Engineering*, 63(7), pp.1455-1462. doi: 10.1109/TBME.2015.2496264.

Vayena, E., Blasimme, A., & Cohen, I. G. (2023). Machine learning in medicine: Addressing ethical challenges. PLoS Medicine, 15(11), e1002689. https://doi.org/10.1371/journal.pmed.1002689.

Wang, Z., et al., 2023. Differentially private generative decomposed adversarial network for vertically partitioned data sharing. Information Sciences, 619, 722-744.

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., ... & Goldenberg, A. (2019). Do no harm: a roadmap for responsible machine learning for health care. Nature Medicine, 25(9), 1337-1340. https://doi.org/10.1038/s41591-019-0548-6.

World Health Organization. (2022). Breast cancer. https://www.who.int/news-room/fact-sheets/detail/breast-cancer.

Xie, T., et al., 2024. Deep Matcher: A Deep Transformer-based Network for Robust and Accurate Local Feature Matching.

Xu, T., et al., Multimodal Deep Learning for Cervical Dysplasia Diagnosis. In: Multimodal Deep Learning for Cervical Dysplasia Diagnosis.

Yaqoob, A., Musheer Aziz, R., & Verma, N. K. (2023). Applications and Techniques of Machine Learning in Cancer Classification: A Systematic Review. Human-Centric Intelligent Systems, 3, 588–615. https://doi.org/10.1007/s44230-023-00041-3.

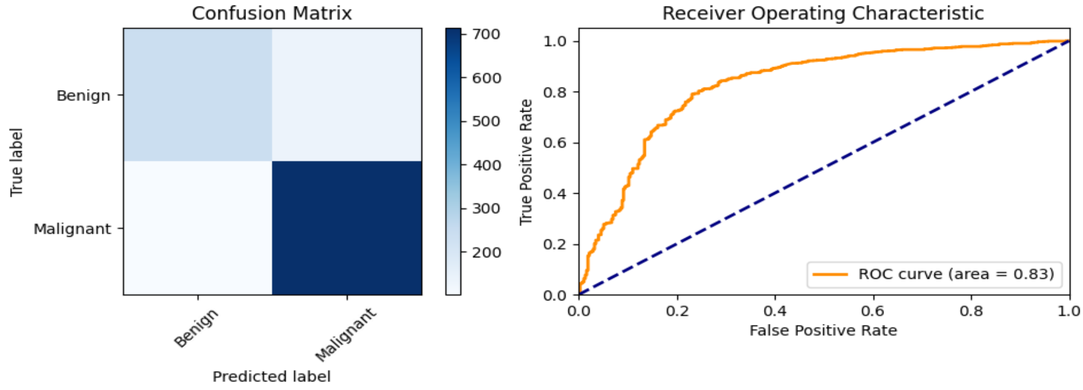
Yaqub, M., Jinchao, F., Aijaz, N., Ahmed, S., Mehmood, A., Jiang, H., & He, L. (2024). Intelligent breast cancer diagnosis with two-stage using mammogram images. *Scientific Reports*. DOI: 10.1038/s41598-024-65926-0.

Zakareya, S., Izadkhah, H. and Karimpour, J., 2023. A New Deep-Learning-Based Model for Breast Cancer Diagnosis from Medical Images. *Diagnostics*, 13(1944), pp.1-23. Available at: https://doi.org/10.3390/diagnostics13111944 [Accessed 29 July 2024].

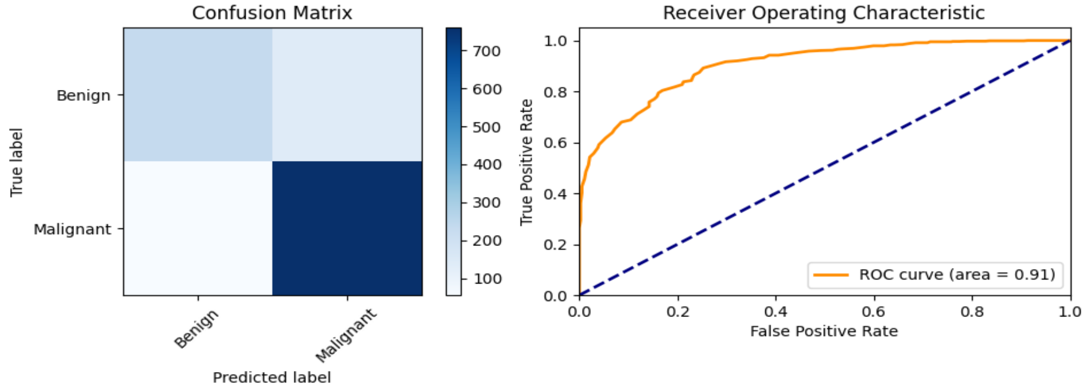
Zhong, Y., et al., 2024. A multi-task fusion model based on a residual–multi-layer perceptron network for mammographic breast cancer screening. Computer Methods and Programs in Biomedicine, 247.

Zouhri, R., Azouazi, M. and Haouachka, N. (2024). Feature Selection Methods in Data Mining for Classification. Recent Advances in Information and Communication Technologies, pp.163-180.

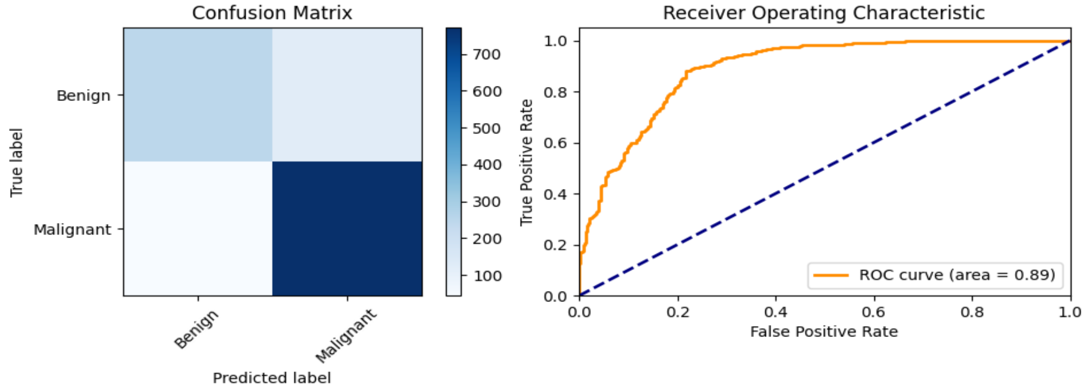
APPENDIX A



Appendix 1:Confusion Matrix and ROC curves for SVM Model without Feature Engineering

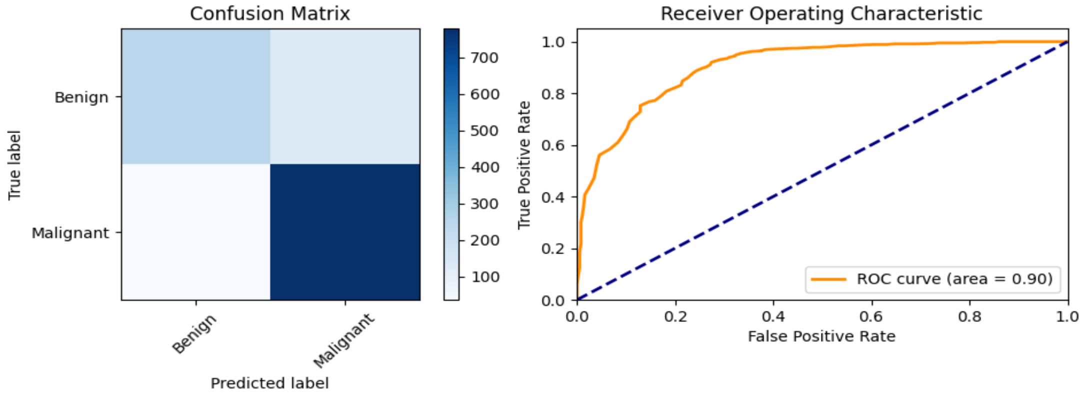


Appendix 2:Confusion Matrix and ROC Curves for Random Forest Model

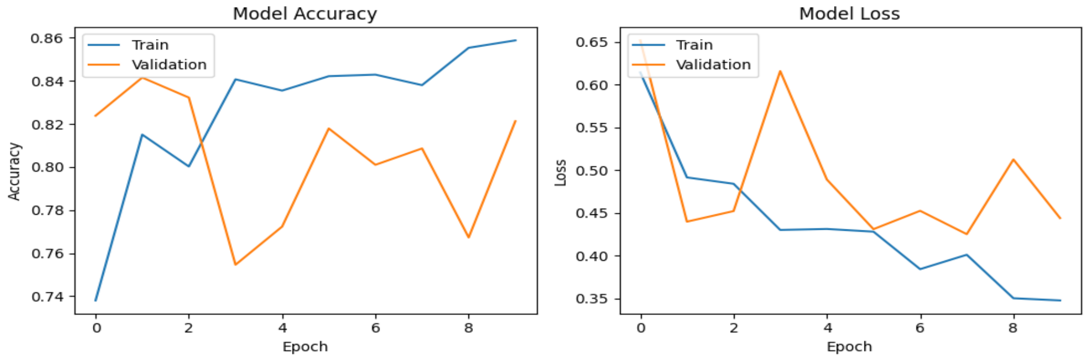


Appendix 3: Confusion Matrix and ROC Curves for SVM with feature engineering.

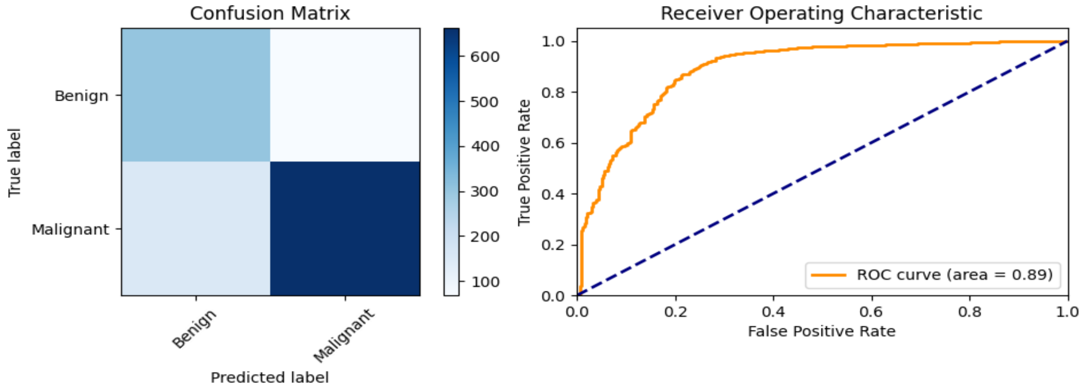
APPENDIX B



Appendix 4: Confusion Matrix and ROC Curves for Random Forest Model with Feature Engineering.

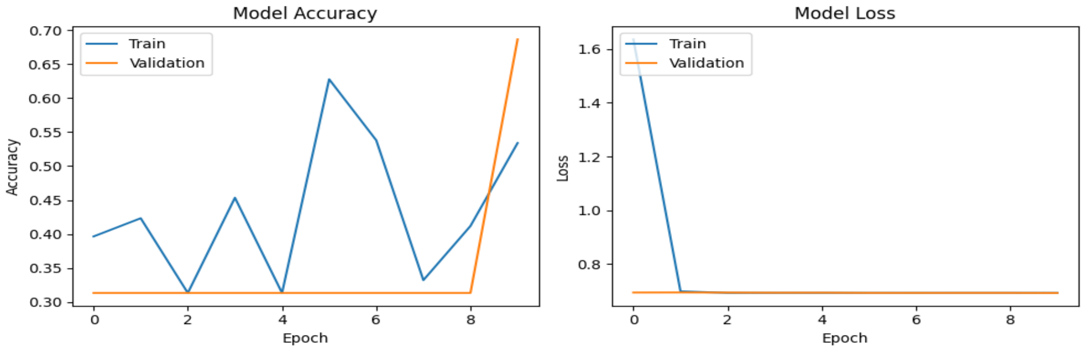


Appendix 5: Training and Validation Accuracy/Loss Curve for Single Modal CNN

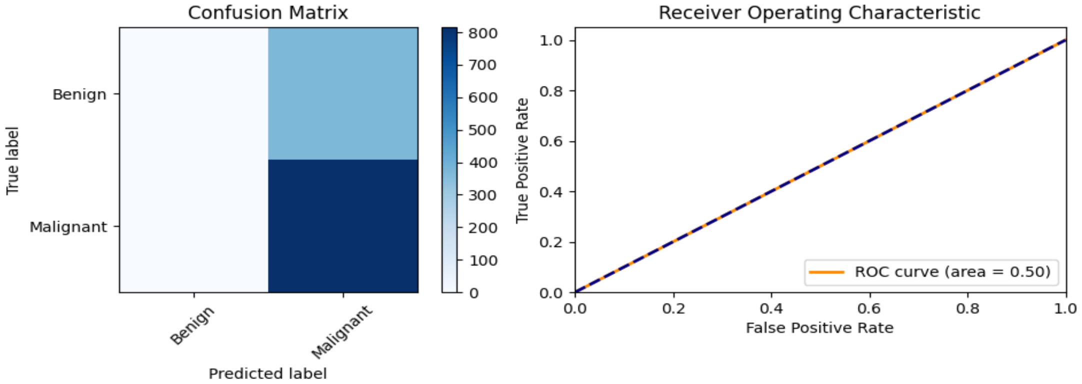


Appendix 6: Confusion Matrix and ROC Curve for Single Modal Custom CNN

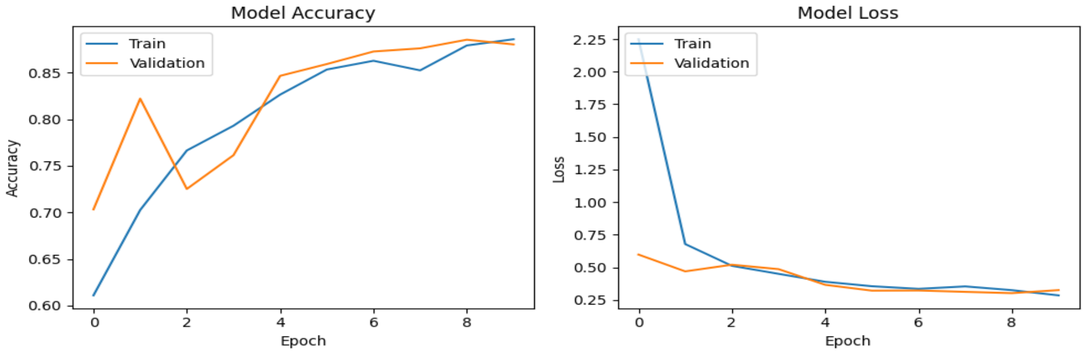
APPENDIX C



Appendix 7: Training and validation Accuracy/Loss Curve for single modal Resnet 50

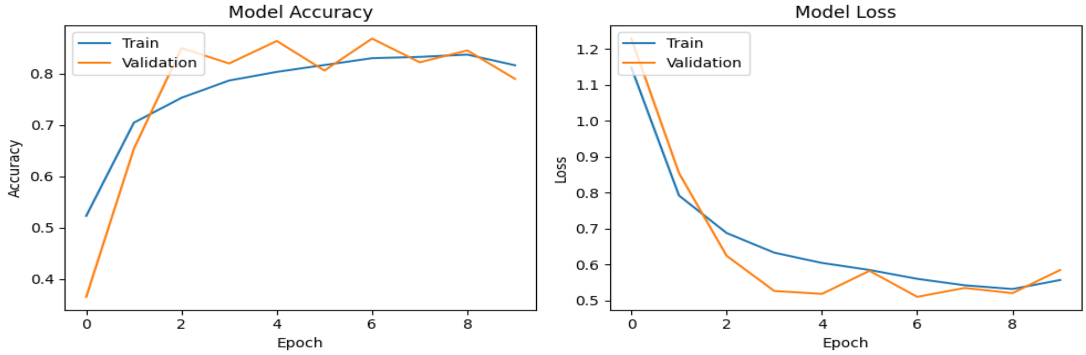


Appendix 8: Confusion Matrix and ROC Curve for single modal Resnet50

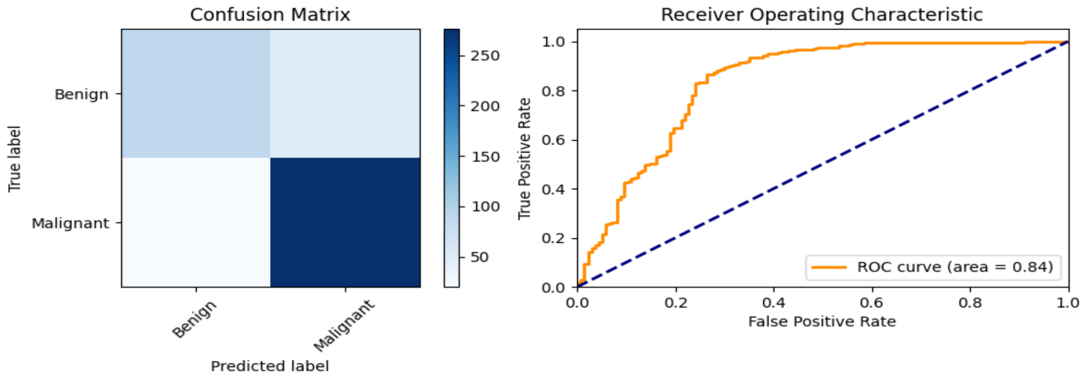


Appendix 9: Training and validation Accuracy/Loss Curve for single modal Densenet121

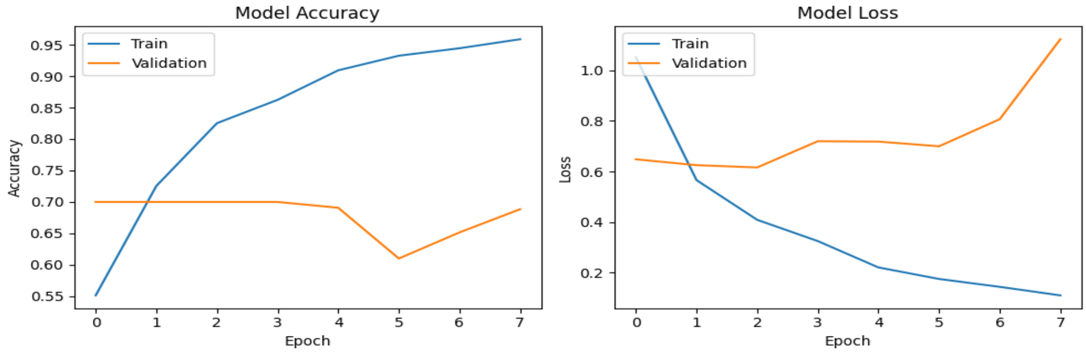
APPENDIX D



Appendix 10: Training and Validation Accuracy/Loss Curves for Dual-Stream Custom CNN

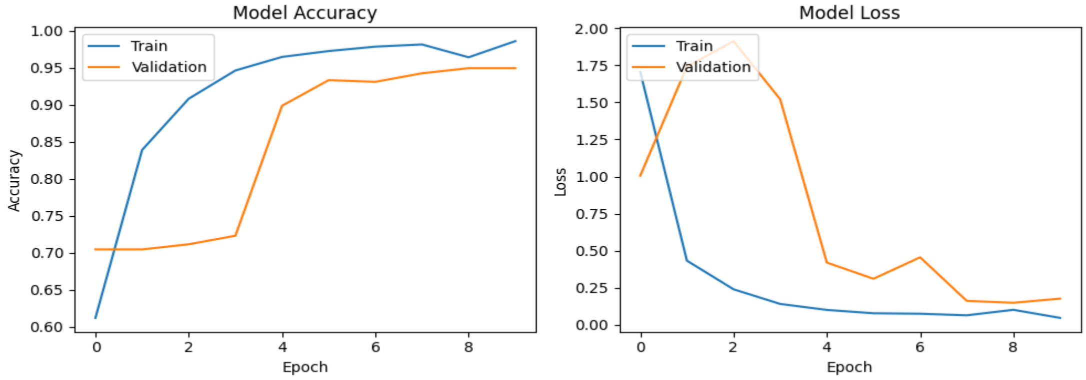


Appendix 11: Confusion Matrix and ROC Curve for Dual-Stream Custom CNN

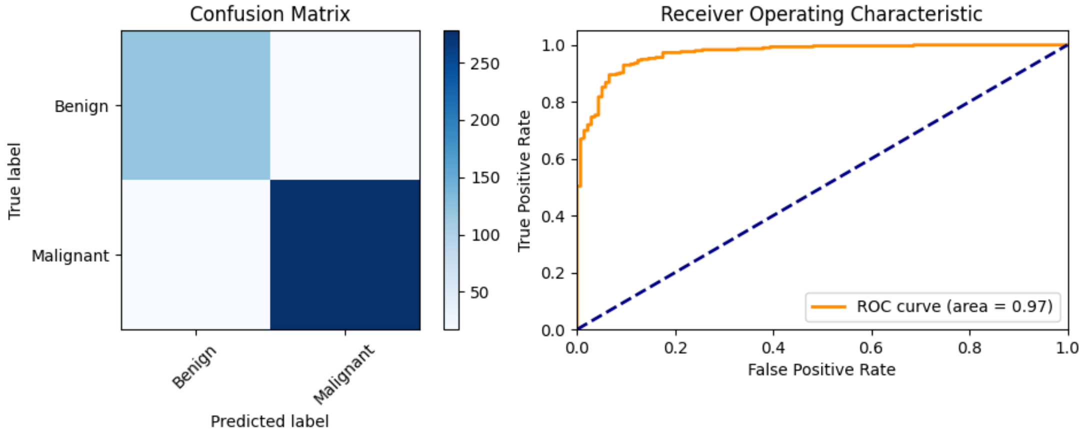


Appendix 12: Training and Validation Accuracy/Loss Curves for Dual-Stream ResNet50

APPENDIX E

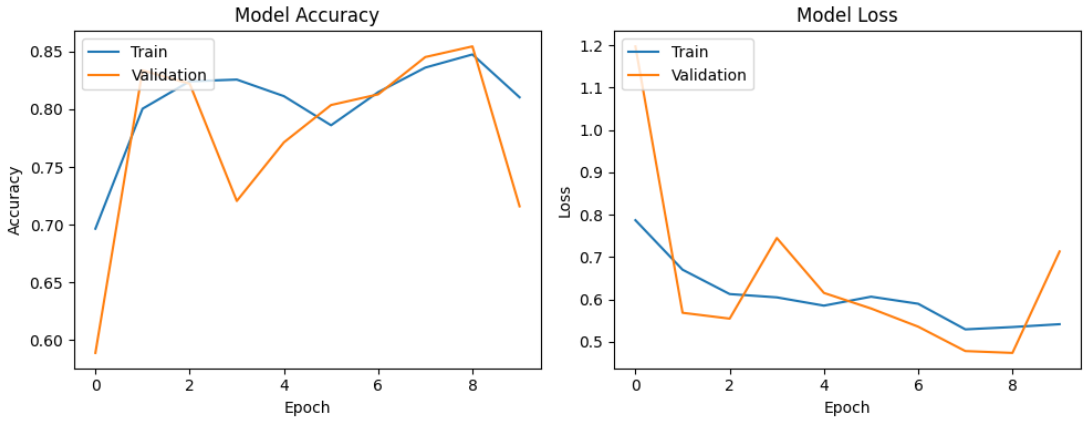


Appendix 13: Training and Validation Accuracy/Loss Curves for Dual-Stream DenseNet121

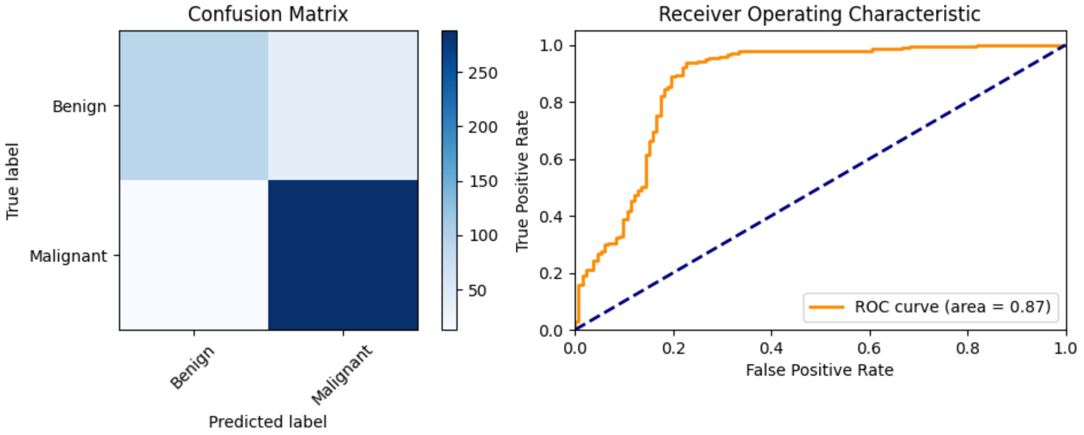


Appendix 14: Confusion Matrices and ROC Curves for Dual-Stream DenseNet121

APPENDIX F



Appendix 15: Training and Validation Accuracy/Loss Curves for dual stream custom CNN with Feature Engineering



Appendix 16: Confusion Matrix and ROC Curve for dual stream custom CNN with Feature Engineering