ANALYSIS OF MEDICAL IMAGES OF BREAST FOR THE PRESENCE OF TUMOURS USING DEEP LEARNING TECHNIQUES

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

SATYANARAYANA REDDY BERAM

MZU REGN NO: 2300002

Ph.D. REGN NO: MZU/Ph.D./2212 of 18.08.2022

DEPARTMENT OF INFORMATION TECHNOLOGY

SCHOOL OF ENGINEERING AND TECHNOLOGY

FEBRUARY, 2026

ANALYSIS OF MEDICAL IMAGES OF BREAST FOR THE PRESENCE OF TUMOURS USING DEEP LEARNING TECHNIQUES

BY

SATYANARAYANA REDDY BERAM

Department of Information Technology

Supervisor: Dr. R. LALCHHANHIMA

Joint Supervisor: Dr. KSH. ROBERT SINGH

Submitted

In partial fulfilment of the requirement of the Degree of Doctor of Philosophy in Information Technology of Mizoram University, Aizawl.

Department of Information Technology
School of Engineering and Technology
MIZORAM UNIVERSITY
(*A Central University*)
Tanhril**,** Aizawl - 796 004, Mizoram

CERTIFICATE

This is to certify that the thesis entitled "**ANALYSIS OF MEDICAL IMAGES OF BREAST FOR THE PRESENCE OF TUMOURS USING DEEP LEARNING TECHNIQUES"** is a record of the research that **Satyanarayana Reddy Beram** completed under our supervision and guidance, and it is submitted to the Mizoram University in the Department of Information Technology under the School of Engineering and Technology, in partial fulfillment of the requirements for the award of the degree of Doctor of Philosophy in Information Technology.

All help received by him from various sources has been duly acknowledged. No part of this thesis has been reproduced elsewhere for the award of any other degree.

Dr. R. LALCHHANHIMA
Supervisor & Associate Professor
Dept. of Information Technology,
Mizoram University, Aizawl-796004

Dr. KSH. ROBERT SINGH
Joint Supervisor, Assistant Professor
Dept. of Electrical Engineering,
Mizoram University, Aizawl-796004

DECLARATION

Mizoram University

February, 2026

I, **Satyanarayana Reddy Beram**, hereby declare that the subject matter of this thesis is the record of work done by me, that the contents of this thesis did not form basis of the award of any previous degree to me or to do the best of my knowledge to anybody else, and that the thesis has not been submitted by me for any research degree in any other University/Institute.

This is being submitted to the Mizoram University for the degree of Doctor of Philosophy in Information Technology.

(SATYANARAYANA REDDY BERAM)

(Dr. R. LALCHHANHIMA)
Supervisor

(Dr. KSH ROBERT SINGH)
Joint Supervisor

(Dr. LALHMINGLIANA)
Head of the Department

Acknowledgement

I am deeply grateful to all those who supported, guided, and believed in me throughout this journey. Their encouragement and contributions have been truly invaluable.

First and foremost, I wish to express my deepest appreciation to my esteemed Ph.D. supervisor, **Dr. R. Lalchhanhima**. His insightful guidance, constructive criticism, and unwavering support have been instrumental in shaping this research. I am deeply indebted to him for his mentorship.

I am equally grateful to my joint Ph.D. supervisor, **Dr. Ksh. Robert Singh**, for his invaluable contributions and steadfast support. His expertise, patience, and relentless encouragement have played a pivotal role in the successful completion of this thesis.

I wish to express my sincere gratitude to **Dr. Lalhmingliana**, Head of the Department of Information Technology at Mizoram University. His visionary leadership and steadfast support have cultivated an intellectually vibrant environment that greatly enriched my research journey.

I am grateful to **Dr. R. Chawngsangpuii**, Associate Professor, Department of Information Technology, Mizoram University, for her constant support, encouragement and planning during the course work.

I thank my doctoral committee members **Dr. V.D.Ambeth Kumar, Dr. Ajoy Kumar Khan**, **Dr. Morrel V.L. Nunsanga** and the faculty members **Mrs. Vanlalmuansangi Khenglawt**, **Dr. Somen Debnath** for their suggestions and comments that have enabled me to complete my research work.

I am sincerely grateful to my father, **Sri B Konda Reddy**, and my mother **Smt. B Nagaratnamma** whose silent sacrifices and unwavering support have been the backbone of my career journey. Their selflessness, strength, and belief in me continue to inspire every step I take.

To my dearest wife, **Smt. B Malleswari**, your love anchored me, your strength protected me. In every silent struggle, your presence spoke louder than words. You walked beside me with grace and quiet courage. This achievement is a reflection of your unwavering heart.

I dedicate this thesis to my son **Mr. B Chetan Saatvic Reddy** for his quiet understanding and unwavering presence that gave me a profound source of strength and motivation throughout the long and demanding journey of this doctoral research.

I extend my thanks to **Dr. S Srilakshmi Parvathi**, Professor, Computer Science and Engineering, KITS, Guntur for her continuous support and suggestions.

I owe my gratitude to my college **KKR and KSR Institute of Technology and Sciences**, Guntur, Andhra Pradesh for providing me the opportunity and support due to which I was able to complete my research work.

Lastly, I would like to convey special thanks to one and all those who have directly or indirectly extended their helping hand in this endeavor.

(SATYANARAYANA REDDY BERAM)

Table of Contents

List of Figures

List of Tables

List of Abbreviations

WHO	World Health Organization
WSIs	Whole-Slide Images
ViTs	Vision Transformers
MRI	Magnetic Resonance Imaging
CAD	Computer-Aided Diagnosis
MIL	Multiple Instance Learning
AI	Artificial Intelligence
CNNs	Convolutional Neural Networks
AUC	Area Under the ROC Curve
FPNs	Feature Pyramid Networks
XAI	Explainable AI
ER	Estrogen Receptor
PR	Progesterone Receptor
CoxPH	Cox Proportional Hazards
C-index	Concordance Index
BCE	Binary Cross-Entropy
GAP	Global Average Pooling
GMP	Global Max Pooling
MLP	Multilayer Perceptron
DDSM	Digital Database for Screening Mammography
DCE	Dice Coefficient
IoU	Intersection over Union
PA	Pixel Accuracy
FPN	Feature Pyramid Network
DANN	Domain-Adversarial Neural Network
GRL	Gradient Reversal Layer
FFDM	Full-Field Digital Mammography
MIAS	Mammographic Image Analysis Society
CV	Cross-Validation
DSC	Dice Coefficient
ACC	Accuracy
PRC	Precision
CI	Confidence Intervals
GT	Ground Truth
SE-CRNN	Spatial–Excitation Convolutional Recurrent Neural Network
BRNNs	Bidirectional Recurrent Neural Networks
SE	Squeeze-and-Excitation
AHE	Adaptive Histogram Equalization
TP	True Positives
FP	False Positives
TN	True Negatives

FN	False Negatives
MSA	Multi-Head Self-Attention
FFN	Feed-Forward Network
QWK	Quadratic Weighted Kappa
CCSF	Cross-Cohort Survival Fusion
PFS	Progression-Free Survival
OS	Overall Survival
HMS-T	Hierarchical Multi-Scale Transformer
MHSA	Multi-Head Self-Attention
ROIs	Regions Of Interest
TCGA-BRCA	The Cancer Genome Atlas – Breast Invasive Carcinoma
AMP	Automatic Mixed Precision
RFS	Recurrence-Free Survival

CHAPTER – 1 INTRODUCTION

1.1. Background and Clinical Motivation

Breast cancer remains the major global health care challenge. According to the reports from the World Health Organization (WHO) and GLOBOCAN, it is now the most frequently diagnosed cancers among women, with more than 2.3 million new cases and approximately 685,000 deaths in 2020 [1]. This burden affects both high-income and low-income countries and underscores the urgent need for reliable early detection strategies. When breast cancer is identified in the early stage, the five-year survival rates can exceed 90%; however, late-stage detection is associated with significantly poor outcomes and more complex and costly treatment pathways. Improving the methods, early and accurate detection are not only a scientific challenge but also a pressing clinical necessity.

The medical image is central to contemporary breast cancer. Mammography remains the primary screening modality used in the population-level programs, enabling the detection of the suspicious masses, micro-calcifications, and architectural distortions. Nevertheless, the performance can be compromised by factors like breast density, heterogeneous tissue and composition, and variable image quality. Concurrently, histopathology based on the whole-slide images (WSIs) continues to serve as the gold standard for definitive diagnosis, staging, and prognostication [2]. For the nuclear morphology and tissue architecture, pathologists assign the stages according to the AJCC criteria [3] and support the downstream treatment planning. Thus, mammography primarily facilitates early case identification, while histopathology provides the conclusive evidence required for the staging and risk stratification.

Despite the clinical importance, these manual interpretative processes face several challenges. In mammography, dense breast cancer tissue may obscure lesions, whereas the overlapping structures and low contrast can hinder mass recognition, contributing to both false negatives and false positives. For the WSIs, the gigapixel scale necessitates prolonged, meticulous examinations, and the diagnostic variability may arise due to fatigue, differing levels of expertise, and the subjective decision making [2]. Moreover, the large-scale screening initiatives generate substantial

volumes of imaging data, increasing the workload pressure on clinicians. Together, these factors are highlighting the need for the computational support systems that enhance the accuracy, consistency, and efficiency throughout the entire diagnostic pathway—from the initial imaging to final staging.

1.2 Limitations of Existing Diagnostic Approaches

Although mammography and histopathology are the central components of breast cancer care, both modalities have inherent limitations that can compromise diagnostic reliability.

Mammography: In women with dense breast cancer tissue, the contrast between the glandular structures and tumour regions is diminished, making small or irregular lesions more difficult to detect and increasing the likelihood they will be overlooked. Differences between both film-based and digital systems in terms of the resolution, noise and overall image quality also influence reader performance and make an algorithm design more difficult.

Histopathology: Manual review of WSIs is time-consuming and can be error-prone. Important grading factors such as mitotic count, nuclear pleomorphism, and glandular differentiation are sensitive to inter-observer variation. This leads to inconsistencies that can directly affect treatment decisions and prognosis. In addition, this kind of manual pipeline does not scale well in busy clinical environments.

Conventional computational approaches: Earlier computational methods tried to reduce some of these problems but could not fully solve them. Hand-crafted features often fail to describe the rich variability of tumour morphology. Early CNN-based systems [4] improved performance compared with classical methods, but they still have several issues:

- **Overfitting and poor generalization** – models trained on one dataset often perform badly on others because of changes in scanning protocol, resolution, or population.
- **Limited receptive fields** – CNNs are strong at local pattern detection but struggle to capture global context over a full mammogram or WSI.

- **Rigid multi-scale fusion** – fixed fusion strategies may wash out subtle but important details.
- **Weakly validated interpretability** – saliency maps and heatmaps are rarely compared systematically with expert annotations, so clinical trust remains low.

These limitations together motivate new approaches that can:

- Generalize across different modalities and cohorts,
- Capture both fine details and global context in an adaptive way,
- Provide interpretability that is checked against clinical ground truth, and
- Scale well enough for real-world deployment.

1.3 Evolution of Deep Learning in Breast Cancer Imaging

Deep Learning has changed the landscape of medical image analysis, especially in detection, segmentation, and staging for breast cancer. The first big progress came from CNN-based architectures, such as U-Net, ResUNet, and DeepLab, which learn hierarchical features directly from image data.

- U-Net uses an encoder–decoder design with skip connections, allowing good boundary segmentation even with limited labelled data.
- ResUNet adds residual connections to deepen the network and improve representational power.
- DeepLab employs dilated convolutions to expand the receptive field and capture broader contextual information.

However, despite these advancements, many models are continuing to struggle with the subtle lesions and irregular tumour morphologies, particularly within the dense breast tissue. This limitation underscores the need for the more selective and robust modelling strategies.

To mitigate some of these challenges, attention mechanisms have been introduced. Spatial attention enables the model to concentrate on the most relevant regions within the image, while channel attention emphasizes the most informative feature maps. Together, these mechanisms enhanced and improve the performance on small,

low-contrast, or irregular tumours and also improve the interpretability by highlighting tumour-related signals while suppressing the background noise.

In parallel, ensemble approaches and multi-task learning frameworks have been proposed to improve robustness and efficiency. Ensembles integrate the strengths of the multiple specialized models, reducing the susceptibility to overfitting on individual datasets. Multi-task learning - whereby related tasks such as segmentation and classification are trained simultaneously promoting shared feature representations and more closely aligns with the real clinical workflows.

A further methodological shift has occurred with the introduction of Vision Transformers (ViTs). These models are leveraging self-attention to capture the long-range dependencies and global contextual patterns, a capability that is particularly advantageous in medical imaging, whereas the diagnostically relevant structures may span large spatial regions and multiple magnification levels [5]. Early results show improvements in the tasks like subtype classification, biomarker prediction, and multi-scale analysis. However, ViTs also bring new challenges, such as high computational cost and the need for stronger interpretability validation.

Since labelled medical data is often scarce and heterogeneous, transfer learning and domain adaptation have become very important. Pre-training on large image datasets (for example, ImageNet) and then fine-tuning on mammography or pathology tasks usually improves performance when annotations are limited. Adversarial domain adaptation and domain-aware normalization further help models to transfer from one dataset to another, for instance, from DDSM [6] to INbreast or MIAS [7].

Altogether, these advances form the core of modern automated breast imaging systems. Still, problems with generalization, interpretability, and deployment remain, and they slow down wide adoption in real clinics.

1.4 Key Research Challenges

Even with rapid progress, deep learning for breast cancer imaging still faces several important challenges that limit its impact and delay clinical translation.

- **Generalization across modalities and datasets:** Many models work well on their original training dataset, but their performance drops when tested on data from other centers with different resolutions, protocols, or patient populations. This instability makes deployment in diverse clinical environments difficult.
- **Sensitivity to small, irregular, or low-contrast tumours:** Current systems often detect obvious lesions but fail on subtle, ill-defined, or low-contrast tumours, especially in dense breasts—exactly the cases where clinical support is most needed.
- **Fragmented treatment of related tasks:** Tasks such as segmentation, classification, and staging are often handled separately. Without a unified framework, models cannot fully share useful features across tasks, leading to extra computation and weaker overall performance.
- **Rigid multi-scale fusion:** Many existing multi-scale methods use fixed, hand-designed fusion rules that cannot adapt to the needs of each individual case. This can weaken or even remove clinically important information.
- **Under-validated interpretability:** Heat maps and saliency maps are more common now, but they are rarely compared in a systematic way with expert-annotated ground truth. Without this validation, explanations remain superficial, and clinicians may not trust them.
- **Loose coupling between staging and prognosis:** In real clinical workflows, staging and prognosis are strongly linked, but in most models, they are treated as separate tasks with no joint optimization. This weakens the potential for end-to-end decision support.
- **Scalability and deployment issues:** Many state-of-the-art models are large and computationally expensive. This makes it hard to integrate them into routine workflows, especially in resource-limited hospitals or high-throughput screening settings.

Taken together, these challenges show the need for new frameworks that can:

- Generalize well across datasets and imaging modalities,

- Use adaptive multi-scale fusion of fine and global features,
- Include clinically validated interpretability, and
- Align multitask learning (detection, staging, and prognosis) with real clinical practice.

Addressing these gaps is essential to move from promising research prototypes to reliable tools that can genuinely support diagnosis, staging, and prognosis in everyday breast cancer care.

1.5 Research Objectives

1.5.1 Overarching Aim

The main aim of this work is to design and develop advanced, interpretable and generalizable deep learning frameworks for comprehensive breast image analysis. This includes segmentation, classification, staging and early prognostic modelling. The thesis tries to reduce the gap between algorithm design and real clinical use by: improving cross-dataset generalization, and producing transparent, trustworthy outputs that can support both radiologists and pathologists in their routine work.

1.5.2 Objectives

To overcome these limitations, this research proposes a set of research objectives aiming at enhancing the accuracy of segmentation and classification in breast cancer imaging are

- To investigate the different deep learning models for their performance towards breast cancer detection and segmentation
- To customize the deep learning architectures to improve the accuracy of breast cancer cells segmentation and classification by capturing the spatial features and temporal features from underlying medical images
- To develop deep attention-based models to predict breast cancer survival outcomes and tumour staging by highlighting relevant regions of interest in medical images.

1.5.3 Research Hypothesis

Breast cancer diagnosis and prognosis can be significantly improved through the integration of advanced deep learning techniques that combine transfer learning, attention mechanisms, multi-scale feature learning, temporal dependency modelling,

and transformer-based architectures. Such integrated frameworks are expected to enhance the accuracy and robustness of tumour detection, segmentation, classification, staging, and risk assessment across diverse mammography and histopathology datasets.

H1: Segmentation and Classification Hypothesis

The incorporation of transfer learning, attention mechanisms, and adaptive feature fusion significantly improves breast tumour segmentation and classification performance, particularly for small, irregular, and low-contrast lesions.

H2: Multi-Task Learning Hypothesis

Joint optimization of segmentation and classification through multi-task learning significantly enhances diagnostic accuracy and model generalization across heterogeneous mammography datasets.

H3: Spatial–Temporal Learning Hypothesis

The integration of spatial feature enhancement and temporal dependency modelling significantly improves the reliability and robustness of breast cancer classification.

H4: Multi-Scale Histopathology Hypothesis

Multi-scale transformer architectures that integrate information from different histopathological magnifications significantly improve breast cancer staging accuracy and interpretability.

H5: Prognostic Assessment Hypothesis

The combination of histopathological image features and clinical metadata significantly improves breast cancer risk stratification and prognosis prediction.

1.5.4 Contributions of This Thesis

This thesis proposes four interrelated frameworks—TL-ESMNet, MsFuNet, SE-CRNN and HMS-T—to support end-to-end automation in breast imaging and pathology. Together they cover detection, segmentation, classification, staging and preliminary risk assessment.

1.5.3.1 TL-ESMNet (Transfer Learning Enhanced Segmentation)

TL-ESMNet mainly targets segmentation and classification under limited annotation conditions:

- **Pretrained encoders**
    - Uses backbones such as ResNet and VGG16 to get strong feature representations, helps when annotated data is small or noisy.

- **Dual attention mechanism (spatial + channel)**
    - Spatial attention highlights where the tumour is,
    - Channel attention highlights what features are most relevant, especially useful for low-contrast or irregular lesions.
- **Dynamic ensemble fusion**
    - Integrates outputs from multiple specialized models, uses an adaptive strategy that depends on tumour characteristics, increases robustness across different lesion types.
- **Improved sensitivity to small and irregular lesions**
    - Addresses one of the most persistent weaknesses of mammography systems, by focusing on small, subtle and irregular tumours that are clinically high risk.

1.5.3.2 MsFuNet (Mammogram Analysis Framework)

The MsFuNet framework focuses on mammogram analysis with unified segmentation + classification:

- **Context-aware FPN**
    - Uses a feature pyramid network designed to be context-aware, supports multi-scale feature extraction for both microcalcifications and large masses.
- **Cross-Attention Fusion**
    - Integrates fine local details with global contextual information,
    - Improves both segmentation accuracy and classification robustness.
- **Domain-agnostic regularization**
    - Uses adversarial training and strong data augmentation to reduce overfitting to a single dataset,
    - Improves generalization across DDSM, INbreast, MIAS and other mammography datasets.

- **Unified multi-task learning**
    - Jointly optimizes segmentation and classification in a single model,
    - Reduces computational overhead and fits better with clinical workflows where both tasks are needed together.

1.5.3.3 SE-CRNN (Spatial–Excitation Convolutional Recurrent Neural Network)

1) Enhancing Spatial Features with SE Blocks: SE-CRNN uses Squeeze-and-Excitation (SE) blocks to adaptively adjust the importance of each feature channel. This channel-wise reweighting helps the model focus on tumour regions instead of irrelevant tissue.

2) Modelling Temporal Dependencies with Bidirectional RNNs: To capture temporal or sequential information, SE-CRNN integrates Bidirectional Recurrent Neural Networks (BRNNs). This temporal context supports more confident and stable classification decisions.

3) Task-Specific Architectures for Segmentation and Classification: SE-CRNN is organized into two main task-oriented branches:

- **SUNet (Segmentation U-Net with SE blocks)**
    - A U-Net based architecture enhanced with SE modules.
    - The SE blocks help SUNet focus on lesion areas and maintain sharper and more reliable contours.
- **TeResNet (Temporal ResNet with BRNN integration)**
    - A classification framework that combines ResNet's strong spatial feature extraction with BRNN-based temporal learning.

Together, SUNet and TeResNet form the full SE-CRNN framework for joint segmentation and classification.

4) Comprehensive Evaluation on Multiple Datasets: To check the generality and robustness of SE-CRNN, extensive experiments are carried out on: CBIS-DDSM, INbreast, and combined or integrated datasets. Performance is compared with several state-of-the-art methods using common metrics such as:

- Dice Similarity Coefficient (DSC) and Intersection over Union (IoU) for segmentation, Accuracy, Sensitivity, and Specificity for classification.

1.5.3.4 HMS-T (Hierarchical Multi-Scale Transformer for Staging)
HMS-T is designed for histopathology-based staging and early prognosis:

- **ViT branches at 5×, 10×, 20× magnifications**
    - 5× branch models tissue-level patterns,
    - 10× branch models glandular and regional structure,
    - 20× branch focuses on nuclear morphology,
    - Together they give a comprehensive multi-scale representation.
- **Cross-scale attention fusion**
    - Integrates features across magnifications using attention, supports holistic and accurate AJCC staging.
- **Quantitatively validated interpretability**
    - Aligns attention maps with pathologist annotations, uses spatial similarity metrics (for example Dice) to quantify agreement.
- **Lightweight risk-stratification module**
    - Combines histology-derived embedding with clinical metadata (e.g., age, ER/PR/HER2 status), estimates survival risk inside one unified pipeline.

1.6 Scope of the Thesis

The scope of this thesis covers many aspects:

(a) Imaging modalities

- **Mammography datasets**: DDSM, INbreast, MIAS [8] are used for detection, segmentation and classification.
- **Histopathology slides**: TCGA-BRCA cohort is used for staging and preliminary prognostic analysis.

This combination covers the full diagnostic spectrum: from screening-level mammography to definitive histopathological diagnosis and staging.

(b) Tasks covered

- **Segmentation** – Delineate tumour boundaries precisely.
- **Classification** – Separate benign and malignant lesions.
- **Staging** – Assign AJCC stage based on histopathology features.
- **Prognosis** – Provide early estimates of survival risk.

Together, these tasks reflect the clinical continuum from early detection to outcome prediction.

(c) Comparative evaluation

- Each proposed framework is compared with state-of-the-art methods [9], using established metrics like Dice, Pixel Accuracy, F1-score, C-index, etc.

Cross-cutting emphases

Throughout the thesis, special focus is placed on:

- Robust cross-dataset generalization,
- Interpretable predictions validated against expert annotations,
- Clinical integration, considering:
    - scalability,
    - computational efficiency, and
    - Work flow compatibility in real hospitals.

1.7 Thesis Organization

This thesis is organized into eight chapters. Chapter 1 introduces the background, motivation, clinical challenges, and research objectives. Chapter 2 reviews related work in breast cancer imaging and deep learning. Chapter 3 presents TL-ESMNet for tumour segmentation using transfer learning and ensemble fusion. Chapter 4

introduces MsFuNet, a joint segmentation-classification model with multi-scale fusion. Chapter 5 proposes SE-CRNN, integrating spatial attention and temporal context. Chapter 6 describes HMST-Lite for early-stage classification from pathology. Chapter 7 details HMS-T, a multi-scale transformer for staging and prognosis. Chapter 8 concludes the thesis with a summary, key findings, and future research directions.

CHAPTER – 2 LITERATURE REVIEW

This chapter presents the comprehensive review of the key literature. This is related to the breast cancer imaging and the application of deep learning techniques for the tumour detection, segmentation, and staging. The goal of this chapter is to give both clinical and technical backgrounds which are necessary to understand the contributions of this thesis. This chapter begins by tracing the evolution of the breast imaging and the computational analysis from the early computer-aided systems. It then examines the main limitations in current work and how these gaps are motivating the frameworks proposed in the following chapters.

2.1 Background on Breast Cancer Imaging and Analysis

2.1.1 Clinical Burden and the Role of Imaging

Breast cancer represents the most frequently diagnosed in women and remains a major cause of cancer-related mortality worldwide [1 and 2]. Every year, millions of the women are diagnosed, and many countries, are currently running the large-scale screening programs. These large-scale screening programs have helped to reduce the mortality rate.

Mammography remains the primary modality for the breast cancer screening, due to its high spatial resolution and the relatively low cost. However, its performance can be affected by various factors such as the dense breast tissue, overlapping structures, and different types of the acquisition artifacts. To improve the sensitivity and specificity, the mammography is frequently combined with the ultrasound imaging and dynamic contrast-enhanced magnetic resonance imaging (MRI) [4]. These combinations are used especially in high-risk of patients or diagnostically complex cases. Histopathological evaluation of tissue or biopsy samples also remains as the final reference standard for the diagnosis and staging. In real time clinical practice, decisions are made by the integrating information from all these modalities. It includes the entire diagnostic pathway from the initial detection to staging, treatment planning, and prognosis assessment.

2.1.2 Evolution of Computer-Aided Diagnosis in Mammography

As medical imaging volumes are increased and radiologists' face heavy workloads, computer-aided diagnosis (CAD) systems are developed to assist in the

lesion detection and image interpretation [7]. Early CAD tools primarily depended on the handcrafted features and the rule-based algorithms. While these systems are sometimes improving the sensitivity, but not frequently generated the high number of false positives and did not generalize well across the different datasets or scanners and imaging platforms.

Later the generations of CAD systems are beginning to include the classical machine-learning techniques, capable of learning decision rules from annotated datasets. Even so, these approaches are still depended heavily on manually designed features and are struggled to capture the complex tumour shapes, textures, and contextual patterns. They are having limited representation of capacity and lack of robustness such models are underscored the need for the more adaptable and the data-driven approaches [9]. This situation opened the doors for deep learning models, which can learn hierarchical features directly from the raw image data and provides a more comprehensive representation of underlying tumour morphology.

2.1.3 From Traditional Image processing to Deep Learning

The shift from traditional feature-engineered pipelines to the deep learning-based methods represents a major change in the breast image analysis. Deep neural network architectures such as U-Net, ResNet, and DenseNet become a popular for segmentation and classification of tasks. These models are popular due to their ability to learn multi-scale representations directly from the imaging data [10 & 11]. In the mammography, these models have been used in mitigating the challenges that are related to low contrast, image noise, and irregular lesion boundaries. In the histopathology, they have been enabled by large-scale, patch-based analysis of whole-slide images (WSIs) at large scale.

However, fully supervised deep learning approaches present several notable challenges. Many methods are required to handle the large numbers of pixel-level or region-level annotations, which are both time-consuming and costly for experts to produce. In addition, working with gigapixel whole-slide images (WSIs) can cause the memory and the computational bottlenecks. To address these limitations, recent researches has investigated the weakly supervised learning, multiple instance learning (MIL), and transformer-based architectures, which are capable of modelling the long-range spatial dependencies. These directions aim to reduce the annotation

effort, better use of global context, and enhance the scalability and interpretability of artificial intelligence (AI) tools in both radiology and pathology workflows [6].

2.2 Deep Learning in Breast Imaging

2.2.1 Early Machine Learning and CNN-Based Approaches

The earliest applications of deep learning studies in the breast imaging mainly used CNNs. U-Net [10] introduced encoder-decoder architecture with skip connections that are rapidly became a benchmark for the medical image segmentation. This architecture enables the generation of precise segmentation masks and preserves fine structures. Its performance may drop, when applied to very large mammograms or images with the substantial tissue overlap. Variants such as SegNet and similar encoder–decoder designs aimed to improve the boundary refinement, while deeper architectures like ResNet and DenseNet further strengthened advanced feature extraction [12].

Despite these methodological advancements, several practical limitations persisted:

- **Computational burden:** High-resolution mammograms and gigapixel WSIs requires a large memory resource and extended the training times.
- **Annotation effort:** The acquisition of reliable pixel-level annotations is costly and slows down the dataset development.
- **Inter-observer variability:** Different experts may annotate the lesions differently, and this subjectivity is transferred into the training labels, making cross-dataset generalization harder.

These challenges have motivated us to design the methods that reduce the annotation requirements, improving the robustness across the imaging centers, and incorporate richer contextual information than what early CNN architectures could capture.

2.2.2 Weakly Supervised Learning

Weakly supervised learning methods aim to reduce the reliance on pixel-level annotations. Within the Multiple Instance Learning (MIL) frameworks, a whole-slide image is treated as a “bag” which is comprising numerous smaller image patches, with only slide-level labels (such as, benign or malignant) provided for supervision

[13]. The model then learns to infer which patches in the bag are most responsible for the overall slide-level classification. More generally, weak supervision leverages coarse labels, heuristics signals, or proxy tasks to guide the model training.

These methods offer improved scalability for large-scale pathology cohorts as they eliminate the need for exhaustive pixel-wise labeling and allow the use of existing diagnostic annotations from the clinical records. Weakly supervised and MIL-based frameworks have demonstrated the promising performance in tasks such as survival prediction and risk stratification [9]. At the same time, they still face many difficulties in accurate lesion localization and robustness to noisy slide-level labels. Despite these limitations, MIL-based frameworks have emerged as practical and widely adopted strategies for WSI analysis, as summarized in Table 2.1.

2.2.3 The Emergence of Vision Transformers in Breast Imaging

Vision Transformers (ViTs) are representing a further paradigm shifted by substituting the traditional operations with self-attention mechanisms. Unlike the convolutional neural networks, which primarily focus on capturing local spatial dependencies, ViTs models are global contextual relationships across the entire image. This property allows them to effectively represent the complex tissue architecture and long-range spatial interaction, which are particularly relevant for both mammography and histopathology.

Recent studies [8] have been reported that ViT models can match or even surpass the CNN baseline models in various breast imaging tasks. For instance, Ayana et al., (2023) [5] reported the high performance in predicting the HER2 status directly from hematoxylin and eosin H&E-stained slides. Furthermore, Hybrid CNN–Transformer models [14] are integrating the convolutional blocks with transformer blocks for capturing the global contextual dependencies. This integration has shown in robust and consistent performance across different magnifications and imaging modalities.

A primary limitation of transformer-based architectures is their requirement for large-scale training datasets and substantial computing resources. To mitigate these challenges, researchers have proposed progressive strategies such as progressive fine-tuning and related training approaches [15] which aim to balance the model accuracy with computational efficiency. As shown in Table 2.1, the methods are tried

to reuse the pre-trained weights, adapted them to new domains, and reduced the overall resource demands which are associated with model training.

Table 2.1: Summary of Key Approaches in Early Deep Learning for Breast Imaging

Approach / Model	Core Contribution	Strengths	Limitations	References
U-Net	Encoder–decoder network for medical image segmentation	Good boundary delineation and fine structure	Performance may drop on very large mammograms	Ronneberger et al. (2015) [10]
SegNet	Pixel-wise segmentation framework	Improved boundary refinement	High memory and computation requirements	Shrivastava & Bharti (2020) [16]
ResNet / DenseNet	Deep CNN families mainly used for classification	Strong hierarchical feature learning	Annotation-intensive; limited long-range context	Hamida et al. (2019) [12]; Huang et al. (2020) [13]
Weakly supervised frameworks	Slide-level or coarse-label learning (e.g., MIL)	Greatly reduces annotation burden; scalable	Less accurate localization; sensitive to noisy labels	Yao et al. (2020) [13]
Vision Transformer (ViT)	Self-attention for global context modelling	Captures long-range dependencies and structures	Needs large datasets and heavy computation	Ayana et al. (2022) [5]
Hybrid CNN–Transformer	Combines local CNN features with ViT global context	Robust across scales and imaging modalities	Fusion design and hyperparameter tuning are complex	Lu et al. (2023) [9]

2.3 Segmentation of Mammograms and Tumour Regions

Accurate segmentation in mammography is a very important step, because it helps to locate the suspicious areas, outline the tumour boundaries, and prepare the clean inputs for later tasks like classification, staging, and risk prediction. Over the years, segmentation methods have moved from simple classical image-processing

techniques to the more advanced deep learning models. This helped to capture the complex shapes, textures, and contextual information in mammograms.

2.3.1 Classical Segmentation Methods

Early segmentation methods [7 and 9] mainly used the basic image-processing ideas such as thresholding, region growing, and morphological operations. Thresholding tries to separate the abnormal tissue from the background by using differences in gray-level intensity. Region growing starts from a seed point and then adds neighboring pixels that look similar, to form a region. Morphological operators are used to clean the segmented regions, reduce noise, and refine object boundaries.

These techniques used in segmentation are quite simple and fast, and they do not need heavy computation. However, they usually perform poorly in real mammograms, where the contrast is low, tumour shapes are irregular, or breast tissue is overlapping. As Shrivastava et al., [16] reported such methods struggle especially when the lesions are small, subtle, or hidden inside the dense tissue. These limitations showed that the need of more powerful, learning-based segmentation approaches in this area.

2.3.2 Deep Learning Architectures for Tumour Segmentation

Deep learning has brought major improvements in mammographic tumour segmentation. Encoder–decoder architectures like U-Net [10] learns context in the encoder path and then reconstructs the pixel wise predictions in the decoder, with skip connections to keep fine spatial details. Variants such as nested U-Net and DeepLabV3+ are further improved the localization by reusing features at multiple levels and by using the aurous (dilated) convolutions to enlarge the receptive field.

In general, mammograms are often very large in size, and processing them directly with deep networks can require a lot of GPU memory and computing time. Multi-scale and hybrid encoder–decoder designs will improve the sensitivity to tumours of different sizes and shapes in processing. In addition, they also increase the number of parameters and make training and fine-tuning more demanding. So, while deep learning gives better segmentation quality, it also brings the practical challenges in terms of computation and optimization.

2.3.3 Attention-Enhanced Segmentation Models

To make the model focus more on clinically important regions, attention mechanisms have been added on top of the standard segmentation networks. For example, Attention U-Net [17] uses the attention gates to highlight lesion-related features and to reduce the background noise. Channel and spatial attention blocks, used in DenseATT-Net [18] adaptively reweight the feature maps so that the informative channels and locations get more emphasis. Cross-attention mechanisms [19] also help to combine the fine local details with broader contextual information.

These attention-based methods usually improve the segmentation accuracy and make the model as more interpretable for clinicians. But they also increase the overall model complexity and may cause to overfitting, especially when the training data are limited or not very diverse.

2.3.4 Transfer Learning in Tumour Segmentation

Transfer learning is widely used to deal with the shortage of annotated breast imaging data. In this case, encoders like ResNet, VGG16, or Inception are fine-tuned for mammography, ultrasound, or MRI segmentation tasks. For example, DeepLabV3+ with a ResNet backbone have been applied to breast ultrasound segmentation and showed the improved robustness in low-contrast images. Leong et al. (2022) [20] reported that the ResNet-50 based transfer learning can accurately segment the micro-calcifications in mammograms, which is important for early diagnosis.

Similar transfer learning benefits have been seen in breast MRI [21] and automated breast ultrasound [22], as summarized in Table 2.2. At the same time, differences in imaging protocols, scanners, and patient populations can reduce that, how well a transferred model works on a new dataset. This motivates the use of the domain adaptation and related techniques to make the segmentation performance more stable across different cohorts and centers.

2.3.5 Multi-Scale Feature Extraction

Breast lesions appear at many different scales, from very small micro-calcifications to the large masses with spiculated margins. Because of this nature, the multi-scale feature extraction becomes essential. Architectures like Feature Pyramid Networks (FPNs) and nested U-Nets will combine the features from several

resolutions, so that both fine local structures and global shape information can be captured [23 & 24] from sources. When a multi-scale attention is added, the network can become more robust to noise and low-quality images [25].

However, if the fusion of multi-scale features is done in a fixed or rigid way, some important details may still be lost or suppressed. Table 2.2 presents the need for adaptive fusion strategies that can dynamically adjust the weights of features at different scales for each individual case, rather than using the same fusion rule for all images.

Table 2.2: Summary of Segmentation Approaches in Mammograms

Approach / Model	**Core Contribution**	**Strengths**	**Limitations**	**References**
Thresholding / Region-growing	Classical intensity-based segmentation	Simple to implement; very fast	Poor in low-contrast images; struggles with irregular tumours	Shrivastava & Bharti (2020) [16]
U-Net / Nested U-Net	Encoder–decoder framework for pixel-wise masks	Good boundary localization; preserves fine details	Not efficient for very large mammograms	Ronneberger et al. (2015) [10]
DeepLabV3+	Uses atrous convolutions for semantic segmentation	Captures multi-scale context	Needs large labelled datasets and strong hardware	Gomez-Flores & Pereira (2020) [4]
DenseNet / EfficientNet	CNN backbones with feature reuse and efficiency	Strong feature learning; faster convergence	High memory use; may need domain adaptation for new data	Huang et al. (2020); Tan & Le (2021) [11]
Attention U-Net / DenseATT-Net	Attention-guided segmentation	Better precision and more interpretable feature focus	Higher computational cost; risk of overfitting	Mridha et al. (2023) [17]; Zhang et al. (2022) [18]
Transfer Learning (ResNet, VGG16)	Fine-tuning pretrained encoders for breast imaging	Reduces annotation needs; works across modalities	Limited generalization across very different datasets	Leong et al. (2022) [20]; Falconí et al. (2020) [6]

FPN / Multiscale Architectures	Multi-level feature extraction and fusion	Sensitive to both small and large lesions	Rigid fusion may drop important details	Peng et al. (2020) [23]; Salama & Aly (2021) [24]

2.4 Generalization and Robustness in Medical Imaging Models

Deep learning systems in medical imaging typically achieve the high performance on the datasets used for training. However, their performance is often reduced when applied to the data from a different hospital, scanner, or patient population. The variations in acquisition protocol, scanner hardware, image resolution, and patient demographics can induce a domain shift. It may substantially reduce the model's accuracy when evaluated outside its original training domain.

2.4.1 Cross-Dataset Variability and Domain Shift

Public mammography datasets such as DDSM, INbreast, MIAS, and CBIS-DDSM exhibits the substantial variability in factors such as image resolution, tissue density representation, and annotation protocols. As a result, a model trained on single dataset frequently demonstrate the reduced the performance when evaluated on a different dataset [26]. For example, a system trained only on film-based DDSM may not generalize well to the digital INbreast images because of the differences in modality and clinical workflow. These dataset-specific biases make the clinical deployment difficult and clearly highlight the need for domain adaptation methods.

2.4.2 Domain Adaptation and Domain-Agnostic Regularization

Many strategies have been proposed to handle the domain shift between datasets. Unsupervised domain adaptation methods try to align the feature distributions of the source and target datasets, even when no labels are available in the target domain [27]. In several works, this idea is implemented using the adversarial training, where the model learns domain-invariant features, so that a domain discriminator cannot easily tell which dataset the features are came from [28].

Inter-cohort variability has also been addressed through different batch normalization schemes that stabilize feature statistics across the centers [29]. Recently, contrastive learning [30] has received the attention because it can learn more robust representations that transfer better across domains. A brief summary of these strategies is given later in Table 2.3.

2.4.3 Regularization and Robustness Strategies

Apart from the domain adaptation, several regularizations and learning strategies are used to improve the robustness:

- **Semi-supervised learning:** Here, the model is trained using both the labelled data and a large pool of unlabeled data. This reduces the dependency on expensive expert annotations.
- **Weak supervision:** In this case, the model is trained using coarse or partial labels instead of dense, pixel-level labels, which is allowing the training on larger cohorts [31].
- **Synthetic augmentation:** GAN-based or heuristic augmentation techniques are used to create the additional samples, especially for minority classes, in order to reduce the class imbalance.

These approaches usually reduce the overfitting and improve the cross-dataset transfer. In addition, they also make the training process more sensitive to hyper parameters and may cause to instability if not tuned carefully.

2.4.4 Limitations for Clinical Deployment

Even with these improvements, some important barriers are still limit the routine clinical use:

- **External validity:** Many models still show a clear performance drop when tested on external datasets that are different from the training data.
- **Resource requirements:** Multi-scale CNNs and transformer-based models are often needed the high-end GPUs and large memory, which may not be available in all hospitals.
- **Interpretability:** Existing explanation methods are sometimes inconsistent and not fully validated, which reduces the clinicians' trust in the system.
- **Evaluation practice:** A large number of studies evaluate their models on a single dataset or only on internal splits, without having the strong cross-cohort validation.

These limitations (see Table 2.3) shown that there is still a strong need for the frameworks that are not only accurate, but also robust, generalizable, interpretable, and practical to deploy in different clinical environments.

Table 2.3: Approaches for Generalization and Robustness in Breast Imaging Models

Strategy	Description	Strengths	Limitations	References
Cross-dataset validation	Testing models across DDSM, INbreast, MIAS, CBIS-DDSM	Shows real-world reliability of the model	Often ignored in many published studies	Gnanasekaran et al. (2020) [26]
Domain adaptation (unsupervised)	Aligns feature distributions across datasets	Reduces dataset-specific bias	Needs careful design; scalability can be limited	Wang et al. (2020) [23]
Adversarial training	Learns domain-invariant feature representations	Improves robustness to domain shift	Increases complexity and may cause instability	Wang et al. (2022) [27]
Batch normalization strategies	Domain-agnostic normalization of features	Simple to use; effective in some settings	May reduce accuracy on very small or noisy datasets	Seo et al. (2019) [29]
Contrastive learning	Learns general representations via similarity constraints	Works well with limited labels	Sensitive to class imbalance and sampling choices	Kim et al. (2021) [30]
Semi/weak supervision	Uses coarse or unlabelled data during training	Lowers annotation burden	May reduce localization precision and fine detail	Liu et al. (2021) [31]
Synthetic data augmentation	Uses GANs or heuristics to generate extra samples	Increases diversity; handles imbalance	Risk of unrealistic artefacts and label noise	Levi et al. (2021) [26]

2.5 Multi-Scale Learning and Feature Fusion

2.5.1 Significance of Multi-Resolution Cues in Histopathology and Mammography

Accurate assessment of the breast cancer inherently requires the examination of imaging data across the multiple spatial scales. Fine-grained features—such as micro-calcifications in mammography or nuclear morphology and texture in histopathology cannot be evaluated individually. They must be interpreted in conjunction with the broader structural patterns, which including the glandular organization and overall tissue architecture.

Sheikh et al. (2021) [32] showed that using both high magnification (around 20× for cellular details) and lower magnifications (around 5×–10× for tissue context) can clearly improves the diagnostic accuracy and the confidence of the reader. A similar idea appears in mammography, where performance improves when the system looks at both small focal lesions and global changes such as asymmetry or architectural distortion.

In total, these findings are making it clear that multi-resolution cues are clinically important for complete tumour detection and proper characterization.

2.5.2 Multi-Scale CNN and Hybrid Architectures

To make use of this multi-scale nature, many researchers have proposed several multi-scale CNNs and hybrid models. For example, Deep-Hipo [33] processes both high- and low-resolution patches so that the network can learn cellular-level information on one hand and the tissue-level context at the same time. In mammography, Feature Pyramid Networks (FPNs) are widely used to extract the features at several scales and to detect the lesions of different sizes [23]. More recent work, for example the multi-view stereoscopic attention networks for ultrasound [22], which combines the information from multiple imaging views to improve the tumour classification.

Hybrid CNN–Transformer architectures [34] are aiming to bring together the strengths of both worlds is: (i) CNN blocks focus on local structural details, while (ii) Transformer blocks capture the long-range global context through self-attention. These developments are important steps towards joining the fine-grained and contextual information inside a single model.

2.5.3 Dynamic and Static Strategies of Feature Fusion

Even though multi-scale models have improved the performance, many of them still use static fusion of features. In such designs, feature maps from different scales are combined with the fixed rules, for example by simple summation or concatenation. This fixed strategy does not adjust the contribution of each scale for a particular case. As a result, small but clinically important signals may be dominated by the strong large-scale patterns, and some subtle findings can be lost.

To overcome this limitation, researchers have started exploring the dynamic fusion methods. MDF-Net [35], for instance, adaptively reweights the multi-scale features depending on the image content for ultrasound tumour segmentation. Adaptive attention fusion [11] assigns different weights to each scale, so that the fusion process can change from one case to another. These content-aware fusion strategies are better suited to handle variability in image quality, anatomical structure, and lesion appearance across patients and datasets.

2.5.4 Identified Challenges

Even with the benefits of multi-scale fusion, some important challenges also remained:

- **Loss of fine-scale cues:** When the features are aggregated in a hierarchical way, very small abnormalities can become diluted or hidden inside the stronger signals.
- **High resource requirements:** Multi-resolution architectures typically demand the substantial memory and computation resources. This can hinder their routine clinical deployment, particularly in resource-constrained environments.
- **Weak linkage across scales:** Many existing models are failing to explicitly integrate the fine-grained local details with the global image context. This limited cross-scale interaction can impair the performance, especially in complex or highly heterogeneous cases as shown in table 2.4.

Collectively, these challenges highlight the continuing need for models capable of adaptively combine local detail and global structure information while maintaining the computational efficiency suitable for practical clinical implementation.

Table 2.4: Summary of Multi-Scale Learning and Fusion Strategies

Approach / Model	Core Contribution	Strengths	Limitations	References
Deep-Hipo (Multi-scale CNN)	Processes both high- and low-magnification patches	Captures cellular-level and tissue-level context	Mostly static fusion; limited flexibility	Kosaraju et al. (2020) [33]
FPN-based methods	Hierarchical multi-scale feature extraction	Detects lesions across a wide range of sizes	Computationally expensive; fusion rules are fixed	Peng et al. (2020) [23]
Multi-view stereoscopic network	Uses multiple ultrasound views for tumour classification	Improves diagnostic accuracy in ultrasound imaging	Modality-specific; relatively high complexity	Ding et al. (2023) [22]
CNN–Transformer hybrids	Combines local CNN detail with global Transformer cues	Balanced representation of local detail and global context	Requires large training data and careful tuning	Khan et al. (2022) [34]
MDF-Net (Dynamic fusion)	Adaptively reweights multi-scale features	Handles heterogeneity; better preservation of fine detail	Higher training complexity and compute cost	Qi et al. (2023) [35]
Adaptive attention fusion	Dynamically assigns importance to each scale	Scale-specific adaptability; case-wise fusion	Sensitive to parameter and training setup	Huang et al. (2022) [11]

2.6 Attention Mechanisms for Breast Imaging

2.6.1 Self-Attention in Vision Transformers

The use of self-attention in Vision Transformers has changed the breast image analysis in a major way. Unlike the normal convolutional networks that mainly look at local neighborhoods, self-attention can relate the pixels or patches that are far away from each other in the image. Because of this reason, the model can capture long-range tissue patterns instead of only the local texture.

Liu et al., [31] showed that such global modelling helps the network to learn coherent tissue-level features, which can improve both classification and segmentation, as summarized in Table 2.5. In histopathology, Ayana et al. (2022) [5] used ViT models to predict the HER2 expression directly from the H&E slides. This result suggests that, in some cases, transformer-based models might reduce the need for additional and costly biomarker tests.

2.6.2 Channel, Spatial, and Cross-Attention Designs

As on, several types of the attention modules that have been proposed to refine the features in mammography and related breast imaging tasks:

- Channel attention gives the higher weight to the most informative feature channels and suppresses the less useful responses [18].
- Spatial attention forces the network to focus more on tumour-related regions, which will usually improve the localization and boundary quality [17].
- Cross-attention helps to balance the fine local details with broader global context, which is especially useful for noisy or low-contrast images [19].

These modules can be used alone or combined in hybrid blocks like CBAM or DenseATT-Net. In most of the studies they give consistent performance gains, because they push the network to pay attention to image regions and channels that are more meaningful from a clinical point of view.

2.6.3 Interpretability Through Attention Maps

Attention is not only useful for accuracy, but also for interpretability. Attention maps shows which parts of the image are influencing the model's decision. These maps can be compared with radiologist or pathologist annotations to check, if the model is "looking" at the right place.

Dabass et al. (2022) [36] reported that attention-based heat maps often overlap well with the tumour regions marked by the experts, which increases confidence in the model behaviors. In addition, other eXplainable AI (XAI) tools such as SHAP and LIME have been combined with the attention to link model decisions. This helps in decision making at diagnostic features like mitotic activity, stromal patterns, or tumour-infiltrating lymphocytes [37].

Human-in-the-loop approaches go one step further, where the experts review and adjust the attention maps. This feedback can help to refine the model and at the same time improve trust, since clinicians see how the system is reasoning [38].

2.6.4 Clinical Trust and Adoption Challenges

Even though attention mechanisms look promising, some practical issues are still remaining before they can be used routinely. Attention maps are intuitive, but sometimes they still highlight irrelevant structures, which may make the clinicians doubt the system. Also, many research articles do not report strong quantitative validation of attention, for example overlap scores (Dice), region-wise precision and recall, or agreement with multiple readers.

Another challenge now is how to bring the attention outputs into daily radiology or pathology workflow. If the interface is not designed properly, the extra heat maps and overlays can increase the cognitive load instead of reducing it. Careful design of displays, alerts, and interaction patterns is needed. To move the attention-based tools closer to real clinical adoption and regulatory approval, it will be important to develop the standard benchmarks for interpretability. This runs the prospective reader studies, and help in designing the workflow-aware interfaces that support clinicians instead of distracting them (see table 2.5).

Table 2.5: Overview of Attention Mechanisms in Breast Imaging

Mechanism / Model	**Core Contribution**	**Strengths**	**Limitations**	**References**
Self-attention (ViT)	Models long-range dependencies	Captures global context; high accuracy	Needs large datasets; high computational cost	Liu et al. (2021) [28]; Ayana et al. (2022) [5]
Channel attention	Re-weights informative channels	Enhances discriminative feature maps	May not fully use spatial information	Zhang et al. (2022) [18]
Spatial attention	Focuses on tumour-relevant regions	Better localization and boundary quality	Can be sensitive to noise	Mridha et al. (2023) [17]
Cross-attention	Links local detail with global context	Robust in low-contrast or noisy images	Higher model complexity; tuning can be difficult	Lu et al. (2022) [19]

Hybrid (CBAM, DenseATT-Net)	Combines channel and spatial attention	Improves both accuracy and interpretability	Extra computational overhead	Zhang et al. (2022) [18]
XAI-integrated attention maps	Aligns attention with clinical annotations	Supports clinician trust and adoption	Validation protocols still not standardized	Maouche et al. (2022) [37]; Dabass et al. (2022) [36]

2.7 Tumour Staging and Prognostic Modelling

Tumour staging and prognostic assessment are playing a central role in breast cancer management. They guide the treatment choices and also help the clinicians and patients to understand the likely outcomes. Traditionally, these decisions mainly depended on careful histopathology review and biomarker scoring. With the growing of the computational pathology and deep learning, more automated pipelines are being developed to improve the reproducibility, scalability, and accuracy.

In this section, we first describe the classical clinical staging systems, then move to the CNN- and transformer-based WSI analysis, and finally discuss about the risk stratification, survival prediction, and multimodal survival models.

2.7.1 Traditional Staging Systems

Nowadays, the AJCC TNM system is the main clinical standard for breast cancer staging. It combines the information on primary tumour size (T), regional lymph node status (N), and distant metastasis (M). In routine practice, this framework is further refined using biomarkers such as HER2, Estrogen Receptor (ER), and Progesterone Receptor (PR) [39].

Although widely accepted and validated, these assessments are still time-consuming and shows the noticeable inter-observer variation. For example, mitotic count shows that only the moderate agreement between pathologists [40] and HER2 scoring can be inconsistent in borderline cases [41]. Such variability can change the final stage in a significant number of patients, which may influence the treatment decisions. These issues motivate the development of semi- or fully automated methods that aim to reduce the subjectivity and streamline the pathology workflow.

2.7.2 CNN- and Transformer-Based WSI Analysis for Staging

The digitization of pathology slides has allowed the deep learning models to be applied at the whole-slide level. Early CNN models such as ResNet and DenseNet achieved the strong patch-level classification and segmentation performance [10]. However, they faced difficulty in modelling the long-range dependencies in gigapixel whole-slide images. MIL addresses part of this problem by treating a WSI as a “bag” of patches and using it only for the slide-level labels for training [13]. This reduces the annotation effort but still has some limitations in localization accuracy.

More recently, ViTs have shown the strong capability in capturing global contextual information through self-attention. Ayana et al. (2022) [5] reported the high accuracy (AUC = 0.92) for predicting HER2 status directly from H&E-stained slides using ViT-based models. Hybrid CNN–Transformer architectures [42] combines the local feature extraction from CNN blocks with global reasoning from transformer layers, which is leading to the improved robustness across different magnifications. Taken together, these developments are indicating that the transformer-based and hybrid models can overcome several scalability and context limitations, which were found in earlier CNN-only pipelines.

2.7.3 Risk Stratification and Survival Prediction

Risk stratification divides the patients into groups with different prognosis, while survival models try to estimate the individual survival curves or hazard over time. Classical tools include Kaplan–Meier survival curves for univariate analysis and Cox proportional hazards (CoxPH) models for multivariate analysis [34]. These approaches are well established, but they rely mainly on manually selected clinical and pathological variables.

Recent works [19, 32] has begun to integrate the deep learning with survival analysis. Deep survival models will replace or augment the hand-crafted covariates with image-based features extracted from histopathology. For example, DiaDeepBreastPRS [43] produces the histology-derived risk scores that are correlate with clinical metadata and improved prognostic performance. In the similar direction, the hist2RNA framework predicts the gene-expression surrogates that

directly form histopathology images, linking the morphological patterns to molecular signals and, ultimately, to survival outcomes.

These methods often achieved higher concordance index (C-index) values than the traditional survival models, which are suggesting that image-driven prognostic modelling has strong potential for precision oncology. A summary of these approaches is given later in Table 2.6.

2.7.4 Multi-Modal Survival Models

Breast cancer biology is complex and depends on multiple interacting factors. Therefore, combining the different data sources can improve prediction quality. Multimodal survival models are integrating the histological features with clinical variables and, when available, genomic or other omics data too.

Wetstein et al. (2021) [44] showed that combining the pathological features with clinical information leads to better risk stratification compared to using each modality alone. Hybrid models that join imaging and omics data have also demonstrated the improved generalization and can sometimes give cost-effective alternatives to the traditional molecular assays. These multimodal approaches move towards more personalized oncology, where imaging and molecular signals are jointly used to estimate the patient prognosis.

2.7.5 Gaps in Linking Staging with Survival Prediction

Even with these advances in survival prediction, several gaps remain in the current literature are:

- **Weak integration between staging and survival tasks:** In many studies, tumour staging and survival prediction are treated as two separate problems. As a result, models do not fully exploit the shared representations that could benefit both tasks and improves the overall prognostic accuracy.
- **Limited use of multi-scale features:** Existing staging frameworks often rely on static fusion strategies, which can struggle to align the fine-grained cellular patterns with large-scale tissue architectures [33 & 34].
- **Incomplete validation of interpretability:** Attention maps or heatmaps are often shown qualitatively, but only rarely evaluated with quantitative metrics

or compared systematically with expert annotations [36]. This makes it hard to judge how the reliable the interpretation actually is.

- **Handling of censored data:** Survival datasets are commonly included with censored cases, where the true event time is not observed. This requires special loss functions and survival-specific modelling. In this case the simple adaptations of standard deep learning losses are usually not sufficient.

These points indicate that future research should focus more on joint staging–survival frameworks, better multi-scale fusion, stronger evaluation of interpretability, and survival models that properly handle censoring see Table 2.6.

Table 2.6: Comparative Review of Approaches for Tumour Staging and Prognostic Modelling

Approach / Study	**Core Methodology**	**Strengths**	**Limitations**	**References**
AJCC TNM + Biomarker Scoring	Manual TNM staging with HER2/ER/PR scoring	Clinically validated; widely used in practice	Inter-observer variability; labor-intensive	Lee et al. (2018) [39]; António et al. (2019) [41]
CNN-based staging models	Patch-level CNN classification / segmentation	Strong local feature learning	Limited modelling of long-range dependencies	Hamida et al. (2019) [12]
MIL frameworks	Bag-of-patches learning with slide-level labels	Reduces annotation burden	Weak lesion localization	Yao et al. (2020) [13]
Vision Transformers (ViT)	Self-attention for global context modelling	Captures long-range dependencies; high accuracy	Needs large datasets; computationally heavy	Ayana et al. (2022) [5]
Hybrid CNN–Transformer	Combines CNN local detail and ViT global context	Robust across scales and magnifications	Fusion often rigid; higher training complexity	Qu et al. (2023) [42]
Kaplan–Meier / CoxPH	Classical statistical survival analysis	Well established; interpretable hazard ratios	Limited to selected variables; no direct image input	Yao et al., (2020) [13]

DiaDeepBreastPRS	Deep histology-derived risk scores	Strong correlation with clinical outcomes	Survival modelling treated separately	Paul et al. (2024) [43]
hist2RNA	Predicts transcriptomics from histopathology images	Links molecular signals to survival	Dual-modality training and integration complexity	Mondol et al. (2023) [45]
Multi-modal survival models	Joint use of histology, genomics, and clinical data	Improved accuracy and generalization	Data harmonization and integration challenges	Wetstein et al. (2021) [44]

2.8 Key Challenges and Research Gaps

Deep learning has led to substantial advances in both the breast imaging and histopathology. However, the translation of these developments into the routine clinical practice remains limited. Several interrelated challenges are persisted, including the dataset bias, model architectures are design, interpretability, and practical deployments. Collectively, these issues delineate the key research gaps that this thesis is addressing.

2.8.1 Limited Generalizability Across Datasets and Modalities

Many models demonstrated the high accuracy when evaluation on the datasets is used for their training. However, the performance was often degraded significantly when applied to the external cohorts like this DDSM, INbreast, or CBIS-DDSM. Such variations are in resolution, scanner hardware, breast tissues, and density, and the annotation protocols create domain shift across the datasets [27 & 5]. As a result, models predictions become less reliable outside the original training domain. These challenges are underscoring the necessity for the domain-adaptive approaches, which is capable of achieving the consistent performance across the different clinical centers and imaging modalities.

2.8.2 Inefficient Handling of Small, Irregular, and Low-Contrast Tumours

Many breast cancer lesions in mammograms images are presenting as faint, poorly delineated, or irregular structures, with the micro-calcifications posing a particular diagnostic challenge. Conventional CNN-based pipelines are frequently fails to

captures these subtle patterns, resulting in false negatives and reduce the sensitivity [16, 20]. Therefore, there is a need of the robust system that must be effectively capture very fine-grained details from image. The system should simultaneously incorporate the broader anatomical contexts, ensuring that the stable performance of both detection and segmentation tasks, even for the challenging a low-contrast cases.

2.8.3 Rigid Multi-Scale Fusion

Recently, the attention maps and XAI models are increasingly integrated into diagnostic models [37]. However, their evaluation in many studies remains the predominantly qualitative. Under the visuals, overlays are present in dataset images. They are not supported by the quantitative comparison with expert annotations [36 and 37]. For this, interpretability must be approached to a more systematic and reproducible manner, which is incorporating the standardized metrics and explicit alignment with this diagnostic ground truths.

2.8.4 Inconsistent Interpretability Validation and Limited Clinical Trust

Although the attention-based visualizations and XAI models are widely reported, their validation still remains inconsistent across the studies. Many studies simply present the heat maps without measuring the overlap with pathologist or radiologist markings [36 & 37]. This makes it difficult for clinicians to judge, whether the model is focusing on the right regions or not. To build the real trust, explanations must be consistent, quantitatively evaluated, and closely aligned with the way experts interpret the images in real life.

2.8.5 Weak Coupling between Staging and Prognostic Modelling

In most of the literature, staging and survival prediction are handled as the separate tasks. Staging pipelines mainly describes the tumour size, nodal status, and extent of disease, while the prognostic models rely on histopathology, genomics, or clinical variables to estimate risk [43 & 45]. In general, these two lines of analysis are rarely integrated in a single framework, where the shared representations are underused and there is limited joint reasoning between the diagnostic and prognostic information. This separation weakens the potential of AI to support for full clinical decision-making from the detection to outcome prediction.

2.8.6 Computational Burden and Deployment Constraints

State-of-the-art architectures, which are especially transformer-based networks and multi-resolution WSI models, often require the large GPU memory and long processing times. This computational burden is a major barrier for hospitals, which are having the limited hardware and can the real-time deployment difficult. In addition, heavy models can be hard to scale to large clinical workloads, which slows downs their adoption in routine practice [42]. Table 2.7 presents the key research challenges and gaps existed in deep learning from breast image analysis.

Table 2.7: Key Challenges and Gaps in Deep Learning for Breast Imaging

Challenge	Description	Impact on Clinical Translation	References
Limited generalizability	Performance drop across datasets and modalities	Reduces reliability and trust in new settings	Wang et al. (2020) [27]; Ayana et al. (2022) [5]
Small / low-contrast tumour detection	Difficulty capturing irregular margins and micro-calcifications	Missed findings and lower sensitivity	Shrivastava & Bharti (2020) [16]; Leong et al. (2022) [20]
Static multi-scale fusion	Fixed feature pyramids that do not adapt to heterogeneity	Loss of fine details; suboptimal segmentation	Peng et al. (2020) [23]; Kosaraju et al. (2020) [33]
Weak interpretability validation	Attention maps rarely evaluated quantitatively	Clinician scepticism; slower adoption	Dabass et al. (2022) [36]; Maouche et al. (2022) [37]
Disjoint staging and survival modelling	Separate pipelines for staging and prognosis	Missed integration of complementary information	Paul et al. (2022) [43]; Mondol et al. (2023) [45]
High computational burden	Memory- and compute-intensive architectures	Limited scalability in real-world deployment	Qu et al. (2023) [42]

2.9 Summary and Motivations for Proposed Frameworks

The whole review in this chapter shows that the deep learning has made strong progress in mammography and histopathology, but several important gaps are still remaining. The main problems are limited cross-dataset generalization, mostly static

and rigid multi-scale fusion, uneven interpretability validation, and some other practical challenges in real clinical deployment.

These observations motivated to design and propose the four main frameworks in this thesis: TL-ESMNet, MsFuNet, SE-CRNN and HMS-T. Each framework is designed to address a set of limitations existed in this problem space, while together they form a more complete response to the challenges identified above.

2.9.1 Motivation for TL-ESMNet

TL-ESMNet focuses on two long-standing issues are: static multi-scale fusion and the difficulty of detecting small, irregular, or low-contrast lesions. Its main components are:

- Transfer learning with pre-trained backbones (such as ResNet-50 and VGG-16) to strengthen feature extraction when labelled data are limited [20].
- Multi-scale attention modules that highlight diagnostically important regions at different resolutions.
- Dynamic fusion using adaptive weighting, so that cross-scale integration is adjusted image by image and can better match the characteristics of each modality or dataset.

2.9.2 Motivation for MsFuNet

MsFuNet is primarily designed to address the challenges of weak generalization and limited contextual reasoning in mammographic analysis. The architecture is built upon CNN foundations and enhanced through the integration of attention mechanisms. Its core design principles include:

- **Context-aware feature pyramids** that allows the network to detect small signals such as micro-calcifications, while also capturing larger masses and architectural distortions.
- **Cross-attention fusion**, which balances the high-resolution detail with broader anatomical context and helps to improve the boundary quality and decision consistency.

- **Domain regularization**, using the adversarial and contrastive strategies to learn domain-invariant features so that the model can transfer better between various datasets like DDSM, INbreast, and CBIS-DDSM.

By combining the segmentation and classification in a single multi-task model, MsFuNet aims to provide robustness across heterogeneous imaging conditions and to reduce redundant computation.

2.9.3 Motivation for SE-CRNN

The main idea of SE-CRNN is to combine spatial attention with temporal modelling in a unified design. The main contributions of the model are:

1) Enhancing Spatial Features with SE Blocks: SE-CRNN uses Squeeze-and-Excitation (SE) blocks to adaptively adjust the importance of each feature channel. This channel-wise reweighting helps the model focus on tumour regions instead of irrelevant tissue. As a result, lesion segmentation becomes more precise and the number of false positives can be reduced.

2) Modelling Temporal Dependencies with Bidirectional RNNs: To capture temporal or sequential information, SE-CRNN integrates Bidirectional Recurrent Neural Networks (BRNNs). This allows the model to combine information across multiple mammographic views or along sequential feature maps, giving a richer understanding of lesion characteristics.

3) Task-Specific Architectures for Segmentation and Classification: SE-CRNN is organized into two main task-oriented branches:

- **SUNet (Segmentation U-Net with SE blocks)**
- **TeResNet (Temporal ResNet with BRNN integration)**

Combined together, SUNet and TeResNet form the full SE-CRNN framework for joint segmentation and classification.

4) Comprehensive Evaluation on Multiple Datasets: To check the generality and robustness of SE-CRNN, extensive experiments are carried out on: CBIS-DDSM, INbreast, and combined or integrated datasets. Performance is compared with several state-of-the-art methods using common metrics such as Dice Similarity Coefficient

(DSC) and Intersection over Union (IoU) for segmentation, Accuracy, Sensitivity, and Specificity for classification.

2.9.4 Motivation for HMS-T

HMS-T is designed to bridge the gap between tumour staging and survival analysis. It uses the hierarchical transformers and multi-scale reasoning to link representation learning with its outcome prediction. The main ideas are:

- **Hierarchical multi-scale transformers** that explicitly model the relationships from cellular-level detail up to larger tissue architecture.
- **Cross-resolution reasoning** which combines the local morphological cues with global context to obtain more robust staging decisions.
- **A unified staging–survival pipeline**, where the lightweight survival head is attached to the staging network so that the prognostic risk is estimated from the same shared representations.
- **Self-attention maps** that support interpretability and can be compared with the expert annotations, while the modular design is meant to ease the integration into clinical workflows as shown Table 2.8.

Table 2.8: Mapping of Challenges to Proposed Framework Solutions

Challenge / Gap	**TL-ESMNet Contribution**	**MsFuNet Contribution**	**SE-CRNN Contribution**	**HMS-T Contribution**
Limited generalizability	Transfer learning to adapt to new datasets	Domain regularization across mammography datasets	Comprehensive evaluation on multiple datasets	Multi-cohort validation in WSI-based staging
Small, irregular, low-contrast tumours	Multi-scale attention with dynamic fusion	Context-aware feature pyramids	Squeeze and Excitation Blocks	Hierarchical cross-resolution reasoning
Static multi-scale fusion	Adaptive fusion inspired by MDF-Net-style strategies	Cross-attention based fusion	Bidirectional Recurrent Neural Networks	Explicit hierarchical transformer-based fusion

Weak interpretability validation	Attention-guided segmentation outputs	Attention-driven feature integration	Deep Feature Maps	Self-attention maps checked against expert feedback
Disjoint staging and prognosis	Strong segmentation across imaging modalities	Focus on detection and classification	Focus on segmentation and classification	Single pipeline for joint staging and survival prediction
High computational burden	Use of lightweight pre-trained backbones	Efficient reuse of features within the network	Combine information along Feature Maps	Progressive fine-tuning for scalable WSI processing

In summary, the TL-ESMNet, MsFuNet, SE-CRNN and HMS-T are proposed as a connected set of frameworks that are extend in previous work and directly address many of the recurring challenges in deep learning for breast imaging. The following chapters describe their architectures, training strategies, and evaluation results in detail.

CHAPTER – 3 ENHANCING BREAST TUMOUR SEGMENTATION WITH TL-ESMNET: A TRANSFER LEARNING AND ENSEMBLE-BASED APPROACH FOR MAMMOGRAMS

3.1 Introduction

3.1.1 Background and Motivation

Breast cancer remains one of the leading causes of the cancer-related mortality among the women. This is accounting for nearly one in four newly diagnosed cancer cases within the female population. Every year, millions of women are affected by breast cancer diagnoses, underscoring the critical importance of early detection and timely therapeutic interventions. When it is recognized at an early stage, the breast cancer is preventable [1].

Mammography is considered as the gold standard modality for the breast cancer screening and it enables the detection of suspicious abnormalities before the clinical symptoms. Mammography diagnostic performance is not always optimal, especially when it is applied in women with dense breast tissue [9 and 25]. In such cases, overlapping the anatomical structures and reducing the image contrast may become ambiguous with underlying lesions [46].

From a computational point of view, the automatic segmentation of breast tumours in mammographic images is still a very challenging problem. Generally, the tumours exhibit large variations in size, shape, borders, appearance, and intensity levels. In addition to this, image noise, low contrast, and acquisition artifacts can further degrade the visual quality of the input images [47]. Conventional segmentation methods, which usually rely on hand-crafted features or fixed parameter settings, often fail to cope with this variability. When these models are trained on one dataset and then evaluated on a different dataset, their performance can typically drop. This is indicating that the limited generalization ability and reducing their suitability for the reliable use in real-time clinical workflows.

Since the past decade, many deep learning-based methods have been designed to enhance the segmentation performance through the feature fusion strategies and more sophisticated network architectures. However, many of these approaches still rely on static fusion mechanisms, in which the contributions of different features representations remain fixed and do not adapt to the specific content of each image. This limitation reduces their ability to effectively capture the small, subtle, or atypical lesions. In addition, the deep models are quite sensitive to the changes in acquisition protocol, resolution, and dataset characteristics [48]. For example, INbreast and DDSM datasets are differ in image resolution, image type (digital vs. film-based), and annotation style. Because of this, a model trained on one of these datasets may not perform well on the others.

To address these challenges, this research proposes a flexible segmentation framework which is called TL-ESMNet. The main ideas behind TL-ESMNet are:

1. **Transfer learning with pretrained models (ResNet, VGG16):** These backbones are used to extract the high-level semantic features, while reducing the need for very large annotated mammography datasets.
2. **Dual attention mechanism (spatial + channel attention):** This design encourages the network to focus on tumour-relevant regions and important feature channels, while suppressing the background noise.
3. **Dynamic fusion strategy inside an ensemble module:** Instead of the fixed fusion weights, the model adaptively adjusts the fusion according to the tumour size, shape, and contrast. This approach is intended to enhance the segmentation accuracy and robustness across a world-wide range of cases.

By integrating the three components, TL-ESMNet aim to address the key limitations which are inherent in the existing methodologies. In specific way, it is designed to operate effectively on both high-resolution digital mammography and lower-quality film-based images, thereby improving its suitability for the real-world clinical settings. The main objective of this framework is to support more reliable diagnosis results, reduce the interpretation errors, and facilitate the earlier intervention in the management of breast cancer.

3.1.2 Objectives and Scope

Building up on the clinical significance and technical challenges that are outlined in above, the primary goal of this work is to develop and evaluate the TL-ESMNet framework. This segmentation framework is designed to achieve high accuracy and strong generalization performance on mammographic images [49]. This research is to structure around the three principal objectives are:

- **High segmentation accuracy:** The framework should be capable to detect and segment the tumours, which are in varying sizes and morphologies. The tumours including small, irregular, and low-contrast masses, in order to reduce the false negatives in real-world screening and diagnostic workflows.
- **Robustness across imaging conditions:** Through the incorporation of transfer learning and adaptive fusion mechanisms, this model is designed to maintain the consistent performance, when applied to the datasets that are different in resolution, noise characteristics, and the breast cancer tissue density.
- **Computational efficiency:** This architecture run within realistic time and memory limits, which is making it possible to integrate into the clinical decision support systems.

To maintain the focus and methodological clarity, the scope of this study is intentionally restricted. All experiments are performed on mammographic images, which are still the primary tool for global breast cancer screening. Within these images, the segmentation task is limited to mass lesions, since both of them are clinically important and visually complex [50]. Other abnormal patterns, such as micro-calcifications or architectural distortions, are not considered in this work due to their micro level. Table 3.1 is summarizing the main challenges in breast tumour segmentation and how the proposed TL-ESMNet framework is addressing them.

Table 3.1: Comparative Analysis of Breast Tumour Segmentation Challenges

Challenge Type	Current Methods	TL-ESMNet Solution
Tumour size and shape variability	Static models do not adapt well to diverse tumour morphology	Dual attention mechanism for focused spatial and channel-wise feature learning

Low image contrast	Lower detection accuracy in dense or noisy breast tissue	Fine-tuned pre-trained backbones to capture subtle and low-contrast patterns
Dataset generalization	Overfitting and poor cross-dataset performance	Transfer learning and dynamic fusion to improve adaptability across datasets
Fusion rigidity	Fixed-weight ensembles underperform in complex or mixed cases	Dynamic fusion that adjusts weights based on tumour properties
Computational constraints	Transformer-heavy models often require high computational resources	Lightweight CNN backbones (e.g., U-Net, DeepLabV3) with carefully integrated attention

In addition, Figure 3.1 illustrates the typical examples of the practical challenges that are motivated this framework, such as differences in tumour size, irregular boundaries, and low contrast. These examples are further support the need for a flexible and adaptive segmentation model.

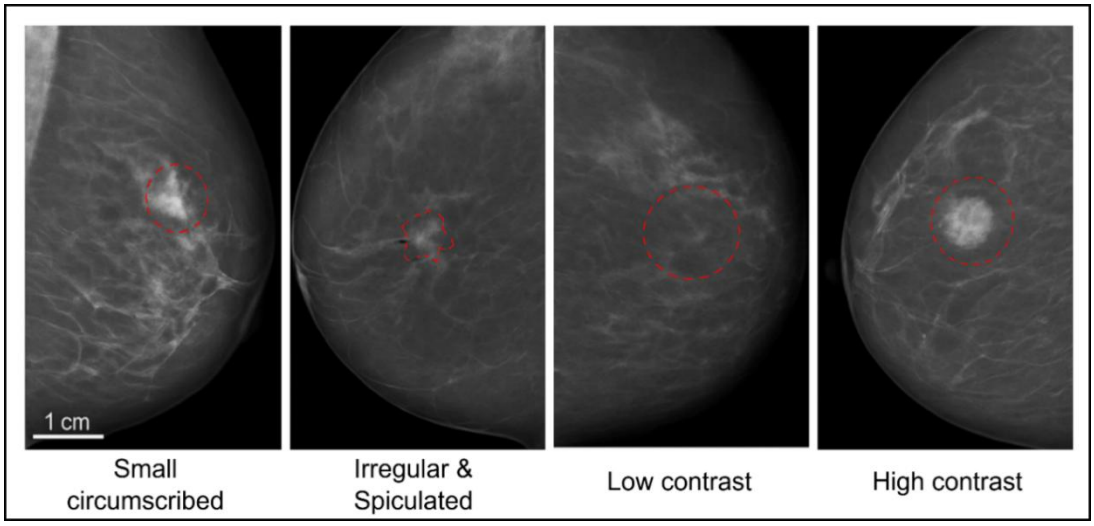

Figure 3.1: Clinical Motivation – Breast Tumour Variability in Mammograms

3.2 Methodology

3.2.1 Overview of TL-ESMNet Architecture

3.2.1.1 Framework Components

The TL-ESMNet framework is designed to address the common challenges that are associated with the breast tumour segmentation by integrating the several custom models within a unified architecture. At the core of this system, a pre-trained feature

extractor is designed based on the CNN models like ResNet-50 and VGG16 [19]. These networks are initially trained on the large-scale ImageNet dataset and are subsequently fine-tuned on mammographic images. This approach enables the models to leverage rich generic feature representations. The simultaneous adaption of these features to the specified visual characteristics in the breast images will improve the segmentation accuracy, even when only a limited number of annotated samples are present in dataset.

To enhance the relation between both tumour and its background regions, the TL-ESMNet incorporates the dual-attention blocks, which are composed with spatial attention mechanisms and channel attention mechanisms. Spatial attention mechanism enables the network to focus on the image regions that are more likely to contain the tumourous tissues. In this manner the channel attention mechanism emphasizes the most informative featured maps and suppressing the less relevant ones. This combined strategy is particularly valuable for the identifying the low-contrast lesions in dense breast tissue [51].

Building upon these attention-enhanced features, the framework employs two parallel segmentation branches: one is U-Net and the other is DeepLabV3. U-Net emphasizes the precise boundary delineation through its encoder–decoder models and Skip connections. On other hand the DeepLabV3 model leverages the multi-scale dilated convolutions to capture the broader structural patterns and the global contextual information.

The outputs of these two segmentation models are subsequently processed by a dynamic fusion module. This module assigns the adaptive weights to the masks generated by U-Net and DeepLabV3 models. The weight depends on tumour-related characteristics such as size, boundary irregularity, and contrast level. As a result, the final segmentation utilizes the fine local details, at the same time, maintains the consistent of global shape also. Figure 3.2 shows the complete TL-ESMNet architecture and how these components interact with each other.

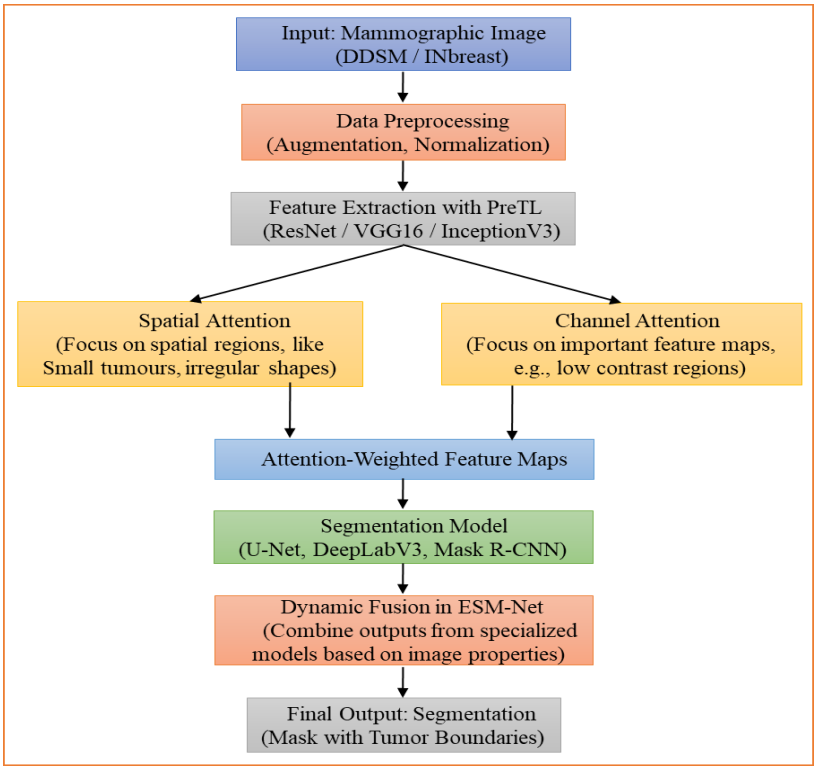

Figure 3.2 TL-ESMNet architecture and its processing pipeline.

3.2.1.2 Workflow Pipeline

The fully processing pipeline of the TL-ESMNet is as follows:

1. **Input preprocessing:** All mammographic images are first resized, normalized, and augmented. Standard operations applied are rotation, horizontal flipping, and brightness adjustment, which help to improve the generalization. For consistency with the backbone networks, each image is resized to 512 × 512 pixels.

2. **Feature extraction:** The preprocessed image is then passed through a pre-trained CNN backbone (ResNet-50 or VGG16) [10 and 14]. This stage produces the hierarchical feature maps that encode structural, textural, and edge-related information relevant for tumour detection.

3. **Application of attention mechanisms:** The extracted feature maps are refined by spatial and channel attention modules. Using the global pooling and learned weighting, these modules emphasize the features that are more indicative of malignancy and reduce the influence of irrelevant background regions.

4. **Parallel segmentation paths:** The attention-weighted features are sent to two parallel segmentation networks are:
    - U-Net, which provides detailed boundary segmentation using its encoder–decoder design with skip connections.
    - DeepLabV3, which captures the multi-scale context and larger tumour structures using aurous (dilated) convolutions.
5. **Fusion and output generation:** Finally, the two predicted masks are combined using a dynamic, weighted fusion as:

$$M_{fused}\,(i,j) \;=\; w_U(i,j)\cdot M_U(i,j) \;+\; w_D(i,j)\;\cdot M_D(i,j) \qquad (3.1),$$

The outputs of U-Net and DeepLabV3 [31] are denoted, and the fusion weights are dynamically computed at each spatial location. These weights satisfy the usual constraint that their sum is equal to one, so the final prediction at each pixel is a convex combination of the two branches. The weights themselves are obtained by a sigmoid-based logistic function as:

$$w_U(i,j) \;=\; \frac{1}{1+e^{-g^{(T_{size}(i,j)-T)}}} \qquad and \qquad w_D(i,j) \;=\; 1 \;-w_U(i,j)\;\; (3.2).$$

This function depends on tumour-related cues such as estimated lesion size, boundary complexity, and contrast. It is controlled by a threshold parameter T and a scaling factor g that adjusts the sharpness of the fusion. The final output is refined by the binary mask that outlines the tumour region with the improved spatial accuracy and consistent global shapes.

3.2.1.3 Key Innovations

Compared to standard segmentation models, our TL-ESMNet introduces three main innovations:

1. **Adaptive dynamic fusion:** Instead of using a fixed-weight ensemble, our TL-ESMNet adjusts the relative contributions of U-Net and DeepLabV3 in an adaptive way based on the characteristics of each image and tumour region [52]. In areas where accurate boundary localization is critical, the framework gives more importance to the U-Net branch. In contrast, for patches dominated by broader, low-contrast structures, the DeepLabV3 branch is assigned a higher weight.

2. **Hybrid attention system:** By integrating the spatial and channel attention mechanisms, the framework is able to parallelly emphasize the locations of salient regions and identify the most informative featured channels. This dual-attention mechanism is especially effective for detecting the subtle or hidden tumours within the dense tissue.
3. **Complementary backbone design:** TL-ESMNet leverages the complementary strengths of U-Net and DeepLabV3. U-Net effectively preserves the fine-grained edges at the micro level, whereas the DeepLabV3 ensures the shape consistency and context at macro level. This complementary interaction is essential to address the heterogeneous tumour images across the datasets.

Collectively, these innovations enabled the TL-ESMNet to generalize effectively across different datasets. These datasets are varying in imaging conditions, and maintaining the robust performance across the different tumour types. It also remains computationally efficient for clinical applications.

3.2.2 Transfer Learning for Feature Extraction

To achieve these robust feature representations across the variability in tumour morphology and imaging quality, our TL-ESMNet is designed on transfer learning strategy. This strategy leverages the models are pre-trained on large extensive imaging datasets. Moreover, these models are subsequently fine-tuned for mammography, rather than training the network totally from the scratch. This methodology facilitates the extraction of semantically rich and broadly generalizable feature before the segmentation stage. All the input images from the DDSM and INbreast datasets are passes through a common preprocessing pipeline. Every gray scale mammogram is resized to 512 × 512 pixels to keep the spatial dimensions consistent between the two datasets. Pixel intensities are then normalized to the range $0, 1$ to support the stable optimization during training. To further improve the robustness, several data augmentation operations are applied on the fly during training process are:

- Random horizontal flip with 50% probability
- Small rotations within ±15°

- Brightness and contrast variation up to ±20%
- Gaussian noise injection with $\sigma = 0.05$

These operations create realistic variations in tumour position, orientation, and visibility, which reflect the heterogeneity observed in clinical practice.

After preprocessing, each image is fed into the chosen pre-trained CNN backbone. The network then converts the image into a set of hierarchical feature maps that contain important cues about the texture, structure, and lesion boundaries [53]. These feature maps are later refined by the attention modules and used as input for the segmentation branches.

3.2.2.1 Model Selection Criteria

Two pre-trained CNN architectures were considered as feature extractors in TL-ESMNet are: ResNet-50 and VGG16. Both are widely used in medical imaging tasks and are originally trained on the ImageNet dataset [57].

- **ResNet-50:** It uses residual skip connections to overcome the vanishing-gradient problem and enables the training of relatively deep networks. It provides robust multi-level feature abstraction and is well suited for capturing the subtle boundaries and irregular tumour shapes in complex backgrounds.
- **VGG16:** VGG16 has a simpler, more uniform architecture made of stacked 3×3 convolutional layers. It is easy to fine-tune for the domain-specific problems and tends to produce dense, low-level feature maps that are helpful for segmentation tasks.

When selecting the most appropriate backbone for mammographic segmentation, the following factors were taken into account:

- The ability to generalize feature representations across different imaging modalities (film-based DDSM vs. digital INbreast),
- The capacity to localize fine-grained tumour structures, and
- The compatibility with the attention modules.

Table 3.1 provides main architectural properties, key characteristics of ResNet-50 and VGG16 as they are used in the TL-ESMNet framework.

3.2.2.2 Fine-Tuning Protocol

To adapt the pre-trained models to mammographic images, a simple layer-wise fine-tuning strategy is used. In both ResNet-50 and VGG16, the early convolution layers are frozen, so their generic low-level filters (edges, basic textures, simple shapes) are kept unchanged. The deeper convolution blocks and the fully connected layers are unfrozen and trained on the target mammography datasets. To reduce the overfitting in these trainable layers, L2 regularization with $\lambda = 0.01$ is applied. Batch normalization layers remain trainable, so that their running statistics can adjust to the intensity distribution and noise characteristics of mammograms. To further stabilize the training and to avoid the strong co-adaptation between neurons, we used the dropout with probability $p = 0.5$ on the unfrozen layers. For optimization, a hybrid loss function is adopted which combines the binary cross-entropy (BCE) and Dice loss [54] with weighting factors $\alpha = 0.7$ and $\beta = 0.3$ as:

$$L_{total} = \alpha \cdot L_{BCE} + \beta \cdot L_{Dice} \quad (3.3)$$

In this way, the model loss balances the pixel-wise classification accuracy (BCE) and region-level overlap quality (Dice). Using this, both the tumour foreground and the background boundaries are optimized in a more stable manner in our model. The detailed fine-tuning settings, including which layers are frozen, which are trainable, and the regularization parameters for each backbone.

Through this carefully controlled transfer-learning process, our TL-ESMNet converts the raw mammograms into high-resolution feature tensors. These tensors preserve the structural, contextual, and morphological information. The enriched feature maps are then passed to the attention modules, which can selectively highlight the tumour-relevant signals before the dual segmentation stage. In this sense, transfer learning acts as the main bridge between the generic pre-trained features and specific characteristics of mammographic imaging. This mechanism supports both the dual-path segmentation and the dynamic fusion mechanisms in TL-ESMNet.

3.2.3 Attention Mechanisms

Accurate tumour classification and segmentation in mammography is often complicated by low contrast, irregular tumour shapes, and noise from overlapping breast structures [55]. To handle these issues, TL-ESMNet employs a dual attention

mechanism, which combines the spatial attention and channel attention. These attention modules operate on the feature maps produced by the pre-trained backbone. They dynamically adjust the feature representation by enhancing tumour-related patterns and suppressing background noise. As a result, the downstream segmentation models will receive more informative inputs and can perform better in difficult cases.

Right after the feature extraction stage, an attention block is inserted. The feature maps from ResNet-50 or VGG16 [42] are processed by the spatial and channel attention units, which recalibrates their spatial layout and channel-wise importance. The refined feature maps are then sent to the two segmentation branches (U-Net and DeepLabV3), so that both local fine details and broader contextual cues are preserved.

3.2.3.1 Spatial Attention

The spatial attention module guides the model about where to look in the mammogram. It encourages the network to focus on the most informative regions, which are typically the areas suspected of containing masses or other abnormalities (see Figure 3.3). This is especially helpful when analyzing the dense breast tissue or lesions with fuzzy boundaries.

The input to this module is a feature map with 'C' channels and spatial dimensions H × W. To construct the spatial attention map, global average pooling and max pooling layers are applied across the channel dimension. These layers produce the two 2D maps that summarize different aspects of the spatial information. These maps are concatenated and passed through a convolution layer with a 7×7 kernel, followed by a sigmoid activation as:

$$M_{spatital} = \sigma\left(f^{7\times7}(Concat[AvgPool(F);\ MaxPool(F)])\right) \quad (3.4)$$

where Conv7×7 denotes the convolution operation, σ is the sigmoid function, and Concat represents channel-wise concatenation. This produces a spatial attention map that encodes both global context and local peaks. The original feature map is then modulated by element-wise multiplication with this attention map:

$$F' = M_{spatial} \odot F \quad (3.5)$$

This step amplifies regions that are more likely to contain a tumour and suppresses less relevant background areas. In this way, the model becomes more sensitive to subtle spatial cues, which improves detection of small or irregularly shaped lesions shows figure 3.3.

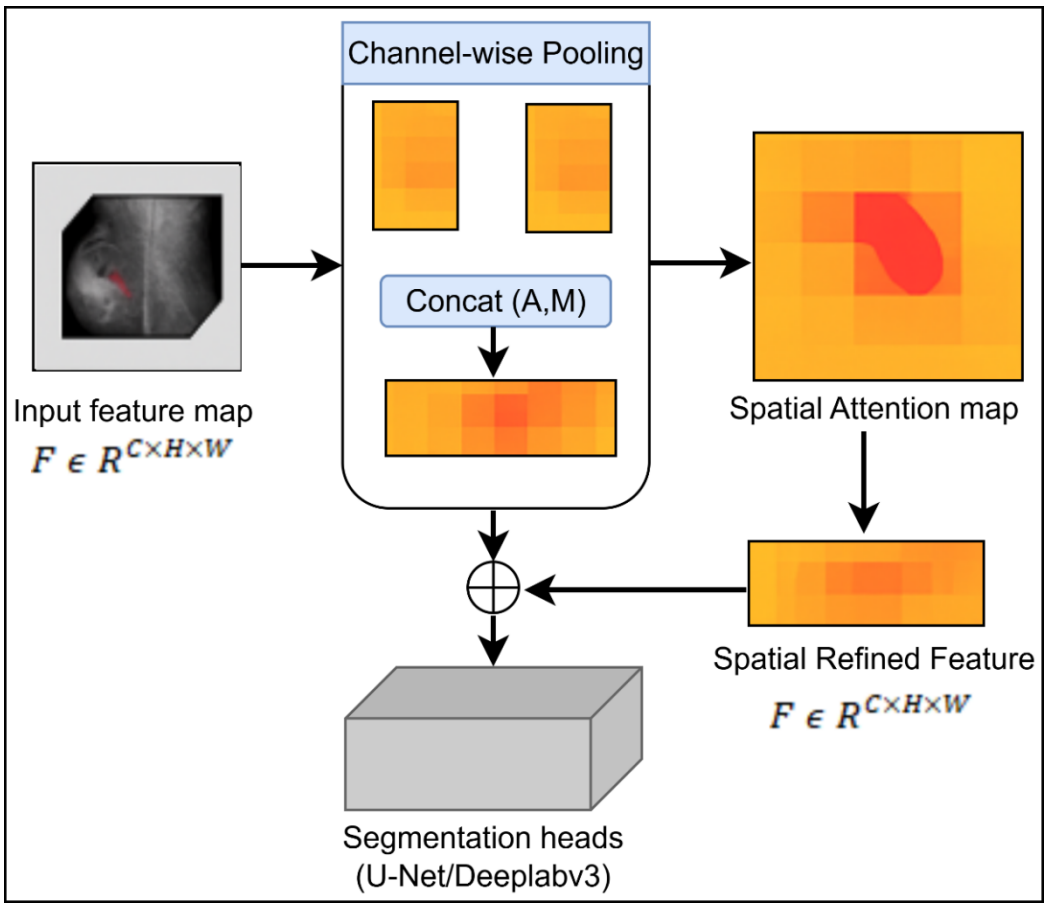

Figure 3.3 Spatial attention process with channel wise pooling, spatial attention and segmentation heads

3.2.3.2 Channel Attention

While spatial attention decides where the model should focus, the channel attention decides what kinds of features are most important. Different channels often encode different semantic information, such as edge orientation, fine textures, or density patterns. However, not all channels contribute equally to tumour detection. Channel attention re-weights these channels so that the most useful channels get higher importance. For an intermediate feature map F, the channel-wise global average pooling (GAP) and global max pooling (GMP) are computed as:

$$z_{avg} = GAP(F), \quad z_{max} = GMP(F) \qquad (3.6)$$

These two descriptors capture complementary information about the distribution of activations across channels. Both are then passed through a shared two-layer MLP with a reduction ratio *r* (typically *r = 16*), ReLU activation, and a final sigmoid:

$$M_{channel} = \sigma\left(W_1\left(\delta\left(W_0 z_{avg}\right)\right) + W_1\left(\delta(W_0\ z_{max})\right)\right) \quad (3.7)$$

Here, the weight matrices are learnable parameters, ReLU is the non-linear activation, and σ is the sigmoid function. The output is a channel attention vector $M_{channel} \in \mathbb{R}^C$, which is then applied to the original feature map as:

$$F'' = M_{channel} \otimes F \quad (3.8)$$

Where $\otimes$ denotes channel-wise multiplication with broadcasting. This operation boosts the channels that are more relevant for tumour appearance and down-weights the redundant or noisy features. When spatial and channel attention are combined, the resulting representation is both spatially precise and semantically focused. These enhanced features are then fed into U-Net and DeepLabV3 [28], which allows the segmentation stage to achieve both accurate localization and strong contextual understanding.

Contribution of Attention in TL-ESMNet

In TL-ESMNet, the dual attention block acts as a selective filtering stage. By modulating the spatial importance and channel relevance in a data-driven manner, it improves the quality of the features that are passed to the segmentation heads. This selective refinement is particularly valuable in challenging mammograms where tumours are partially hidden by dense fibroglandular tissue or appear with very low contrast. These attention modules guide the model where to look and what to emphasize. This helps to delineate the tumours that may be missed or badly segmented by a plain CNN.

3.2.4 Segmentation Backbones

After attention-based refinement, TL-ESMNet sends the enriched feature maps into two parallel segmentation backbones are: U-Net and DeepLabV3. These two architectures were chosen because they provide complementary strengths:

- U-Net is strong in recovering spatial detail and preserving fine edges through skip connections.

- DeepLabV3 captures large-scale contextual information using atrous (dilated) convolutions, which will increase the receptive field without losing resolution.

By combining these two branches, the framework can handle a wide range of lesion sizes, shapes, and imaging conditions, and produces more stable segmentation results.

3.2.4.1 U-Net Implementation

U-Net [10] is a classic encoder–decoder convolutional network that is widely used in biomedical segmentation because it preserves the spatial information across multiple scales. In TL-ESMNet, the attention-weighted feature maps (after spatial and channel attention) are forwarded into the U-Net path [16]. In the encoder, U-Net gradually extracts the higher-level features using convolution layers followed by max pooling. At each stage, the representation becomes more abstract while the spatial resolution decreases. The encoded feature at layer l can be written as:

$$f_e^{(l)} = \sigma\left(W_e^{(l)} * f_e^{(l-1)} + b_e^{(l)}\right) \qquad (3.9)$$

where E^l denotes the encoded feature at layer l, and the weights and biases are learnable parameters with a non-linear activation (typically ReLU). The decoder mirrors the encoder here. It uses the upsampling (or transposed convolutions) followed by convolutions to reconstruct a high-resolution feature map. Crucially, the skip connections link each encoder level with its corresponding decoder level, so that the fine-grained spatial details are re-introduced into the decoding path as:

$$f_d^{(l)} = Upsample\left(f_d^{(l+1)}\right) + f_e^{(l)} \qquad (3.10)$$

These skip connections help to preserve the contours of small and irregularly shaped tumours, which is essential for boundary-sensitive segmentation. Within TL-ESMNet, the U-Net branch mainly focuses on precise anatomical delineation. Its output provides segmentation masks with sharp boundaries and good shape fidelity, making it especially useful for the early-stage or small tumours where exact margins are clinically important.

3.2.4.2 DeepLabV3 Configuration

To balance the fine-detail strength of U-Net, TL-ESMNet also uses DeepLabV3 as the second segmentation backbone. DeepLabV3 is a modern semantic segmentation model that uses an Atrous Spatial Pyramid Pooling (ASPP) block [31] to capture the information at multiple scales. In our framework, the attention-refined feature maps (from the dual attention module) are given as input to the encoder part of DeepLabV3. Inside the ASPP block, aurous convolutions with different dilation rates are applied in parallel as:

$$f_{aspp} = \oplus_{r \epsilon R} AtrousConv(f, r) \qquad (3.11)$$

Where $R = \{1,6,12,18\}$ represents the set of dilation rates, and $\oplus$ denotes the concatenation of feature maps. In this way, DeepLabV3 collects features from several receptive fields at the same time, which helps the model to represent tumours that appear with different sizes and shapes.

$$f_{dlv3} = \sigma\left(W_{proj} * f_{aspp} + b_{proj}\right) \quad (3.12)$$

After the ASPP module, an activation function is used to reduce the feature dimension and refine the representation, while still keeping the important contextual information. The final DeepLabV3 feature map, denoted as f_{dlv3}, contains both coarse global context and useful local detail. This makes DeepLabV3 more suitable for identifying the large, diffuse, or low-contrast tumours that may be missed when the receptive field is too small.

Contribution and Transformation Logic

In TL-ESMNet, U-Net and DeepLabV3 share the same attention-enhanced input features, but they specialize in slightly different roles [21]:

- U-Net focuses on *small and complex tumour contours*, giving sharp boundaries and precise shapes.
- DeepLabV3 focuses more on *global structure and spatial coherence*, thanks to its ASPP and large effective receptive field.

These two outputs act as pre-fusion representations, which are later combined by the dynamic fusion module. This design ensures that both fine textural details and

global contextual cues are preserved in the final segmentation, so the model can adapt to a wide variety of clinical imaging conditions.

3.2.5 Dynamic Fusion Mechanism

Even though U-Net and DeepLabV3 can each produce their own tumour masks, using them independently does not fully exploit their complementary strengths. A simple averaging of their outputs may even reduce performance in challenging cases. To address this issue, TL-ESMNet introduces a dynamic fusion mechanism. Instead of fixed weights, the model learns to adaptively decide how much to trust each backbone based on tumour characteristics [49]. This makes the final segmentation more flexible and better suited to the morphological diversity typically seen in mammographic images.

3.2.5.1 Weight Determination

The fusion process starts by analyzing the attention-weighted features and the initial probability maps produced by U-Net and DeepLabV3. The density and strength of the attention responses give indirect information about the tumour, such as its size, boundary complexity, and contrast [56]. These cues are then used to compute dynamic fusion weights. Let M_{UNet} and M_{DLV3} denote the soft probability masks from U-Net and DeepLabV3, respectively. The final fused probability map M_{fused} is obtained as:

$$M_{fused} = w_{UNet} \cdot M_{UNet} + w_{DLV3} \cdot M_{DLV3}, \quad where\ w_{UNet} + W_{DLV3} = 1 \tag{3.13}$$

Where w_{UNet} and w_{DLV3} are the adaptive weights for U-Net and DeepLabV3, with the constraint that $w_{UNet} + w_{DLV3} = 1$ at each spatial location. These weights are not constant, but they are generated using a sigmoid-based confidence function that depends on an estimated tumour size s as:

$$w_{UNet} = \frac{1}{1+e^{-\alpha(s_t-\mu)}}, \quad w_{dlv3} = 1 - w_{UNet} \tag{3.14}$$

Here, s is a scalar estimate derived from the attention-based activations, α is a sensitivity parameter, and μ is a size threshold that controls the preference between fine-detail segmentation (favoring U-Net) and global-context segmentation (favoring DeepLabV3).

- For small or low-contrast tumours, the function will produce a higherw_{UNet}, so U-Net has more influence.
- For large or diffuse tumours, the weight shifts towards DeepLabV3 (higher w_{DLV3}).

3.2.5.2 Output Integration

Once the adaptive weights are computed, the fusion module integrates the two soft masks. Both U-Net and DeepLabV3 outputs are treated as pixel-wise tumour likelihood maps. The fusion step combines these maps according to their respective weights, so that:

- the high-resolution boundary information from U-Net is preserved, and
- the global consistency and context from DeepLabV3 are maintained [57].

After this weighted combination, a binary tumour mask is obtained by applying a threshold (i.e., 0.5) on the fused probability map. When needed for clinical use, simple post-processing steps such as morphological opening/closing or hole filling can be applied to clean the final segmentation. In this way, the fusion here is not static, but context-sensitive. It adapts to the specific presentation of each tumour, instead of using the same mixing rule for every case. As a result, TL-ESMNet can produce more reliable and generalizable segmentation outputs across mammograms with different resolutions, image quality levels, and tumour morphologies.

At glance, the dynamic fusion mechanism represents a core contribution of TL-ESMNet. By modulating the balance between U-Net and DeepLabV3 according to the tumour features, the framework significantly improves segmentation robustness and makes the model more suitable for real clinical environments [23].

3.3 Experimental Setup

3.3.1 Datasets Used for Evaluation

To evaluate the TL-ESMNet framework, we use two publicly available mammography datasets are: Digital Database for Screening Mammography (DDSM) and the INbreast datasets [58]. These two datasets are selected because they are complementing to each other. DDSM dataset mainly contains a low-quality, film-based mammograms with more noise and artifacts. On other hand, the INbreast

dataset offers the high-resolution and digital images with cleaner visualization. By testing the TL-ESMNet on both DDSM and INbreast datasets, we can check its behaviours of the datasets under the different acquisition protocols, image qualities, and lesion types [6]. In simple words, the DDSM dataset tests the model in difficult conditions, and INbreast dataset tests it in ideal conditions.

DDSM Dataset: The DDSM dataset contains more than 10,000 digital mammograms from around 2,620 screening cases. The original films were scanned at a resolution of about 43.5 microns per pixel. The dataset covers a wide range of breast densities and includes both masses and calcifications. For many images, pixel-level segmentation masks and diagnostic metadata are available. DDSM is especially useful because it reflects older clinical imaging conditions. The images often show: non-uniform intensity distributions, Scanning artifacts, and Annotation inconsistencies. These factors make the dataset quite challenging, but also very realistic for older or resource-limited environments. In this work, DDSM is mainly used to check the robustness of TL-ESMNet in noisy and low-contrast images. The dual segmentation backbones (U-Net and DeepLabV3) and the dynamic fusion mechanism are tested on cases where mass boundaries are weak and tissue overlap is high. This helps to see how well the framework stabilizes feature extraction in difficult scenarios.

INbreast Dataset: In contrast, the INbreast dataset consists of 410 full-field digital mammograms from 115 cases. These images were acquired using modern digital mammography systems and are stored in DICOM format with structured metadata [59]. Pixel-wise annotations are provided for masses, micro-calcifications, and architectural distortions, and they have been validated by expert radiologists. The INbreast images [67] have High and consistent resolution (up to 3328 × 4084 pixels), Low noise levels, and uniform illumination. Because of this, they are well suited for evaluating how precisely the TL-ESMNet can follow tumour margins and fine boundaries. In this thesis, INbreast is treated as a kind of “upper-limit” benchmark, where imaging quality is good and artifacts are minimal. On this dataset, the focus is to test:

- The attention mechanisms (spatial and channel), and
- The adaptive fusion strategy,

To see how well they maintain segmentation accuracy for small, irregular, or low-contrast tumours when image quality is not a limiting factor. Table 3.2 gives a side-by-side view of both datasets and shows how they support a balanced evaluation under both challenging (DDSM) and clean (INbreast) imaging conditions.

Table 3.2: Dataset Statistics and Characteristics

Dataset	**Resolution (pixels)**	**# Cases**	**# Images**	**Lesion Types**	**Annotation Type**	**Imaging Type**
DDSM [6]	~3000 × 5000 (scanned film)	2,620	10,000+	Masses, Calcifications	Pixel-level masks	Film (digitized)
INbreast [67]	Up to 3328 × 4084	115	410	Masses, Micro-calcifications	Pixel-level masks	Digital FFDM

3.3.2 Data Preprocessing Pipeline

Medical image segmentation often suffers from limited annotated samples, strong inter-patient variability, and differences in acquisition settings between centers. To reduce these problems and to help TL-ESMNet to generalize better, we implemented a common preprocessing pipeline for both DDSM and INbreast.

This pipeline has two main parts:

1. Data augmentation, to increase variability and reduce overfitting.
2. Normalization and resizing, to bring all images into a consistent format.

3.3.2.1 Data Augmentation Strategies

In both datasets, we face:

- Class imbalance (e.g., fewer malignant cases),
- Image homogeneity (similar background and breast tissue patterns), and
- Acquisition bias (fixed scanner protocols, fixed view patterns) [60].

To address these issues, several augmentation techniques were applied on-the-fly during training. The main strategies are:

- **Horizontal flipping:** This makes use of the approximate bilateral symmetry of the breasts. By mirroring the images, the effective training set size is increased, and the model sees tumour patterns in both left and right orientations.
- **Rotation (±15°):** Random rotations within a range of about ±15 degrees simulate small changes in patient positioning and tube angle [22]. This helps the model become less sensitive to exact orientation.
- **Contrast jittering (±20%):** Random changes in contrast (up to ±20%) imitate differences in illumination, exposure settings, and scanner calibration between centers. This reduces overfitting to a particular institutional style.
- **Gaussian noise injection (σ = 0.05):** Adding mild Gaussian noise approximates artifacts related to digitization and sensor noise, especially relevant for film-based data such as DDSM. This makes the model more robust to noisy backgrounds.

Together, these augmentations expose the network to a wide range of realistic variations in appearance, intensity, and orientation, while still preserving the basic anatomy and tumour structure.

3.3.2.2 Image Normalization and Resizing

After augmentation, every mammogram and its corresponding ground truth mask were processed to match the input requirements of TL-ESMNet:

- **Resizing:** All images and masks were down-sampled to 512 × 512 pixels. This size was chosen to keep the enough detail for tumour localization (including small lesions) but still allow feasible training on available GPU memory.
    - For images, bilinear interpolation was used.
    - For binary masks, nearest-neighbor interpolation was applied [61].
- **Gray-level normalization:** Pixel intensities were normalized using min–max scaling to the range [0, 1]. This step reduces bias due to differences in scanner hardware, bit depth, or acquisition environment. It makes the inputs

more compatible with the pre-trained backbones (ResNet and VGG16), which were originally trained on ImageNet.

These preprocessing steps ensure that the networks receive the inputs that are both clinically meaningful and numerically stable inputs.

3.3.3 Training Configuration and Environment

Designing an effective training setup is a very critical role for medical imaging segmentation. It focuses on the scenarios where the datasets are imbalanced whereas lesions are occupying only a small portion of the full image. For the TL-ESMNet, the training configuration was precisely selected to ensure:

- Stable convergence,
- Good generalization, and
- Practically useful for segmentation accuracy.

3.3.3.1 Loss Function Composition

In mammograms, tumour regions are occupying only a small portion of the full image. This results in a class imbalance between the foreground (tumour) and background. Utilizing a standard pixel-wise loss function, such as binary cross-entropy (BCE) in isolation tends to bias the models toward the dominant background regions.

To address these challenges, TL-ESMNet employs the composite loss function that integrates the Dice Loss with BCE [62]:

- Dice Loss (L_Dice) focuses on the overlap between the predicted tumour mask and the ground truth. It gives a stronger gradient region to smaller regions are helps to preserve the tumour contours.
- Binary Cross-Entropy Loss (L_BCE) looks at pixel-wise classification quality and encourages stable probability estimates across the whole image.

The total loss with weighting factors $\alpha = 0.7$ and $\beta = 0.3$ is defined as:

$$L_{\text{total}} = \alpha L_{\text{Dice}} + \beta L_{\text{BCE}} \qquad (3.15)$$

This approach places the slightly greater emphasis on the Dice loss function to improve the segmentation of small lesions and the boundary accuracy. On other hand

the BCE components are continuing to ensure the total pixel-wise consistency. The relative weights of this loss terms were determined by experiments and consistently based on high robust performance on both INbreast and DDSM datasets.

3.3.3.2 Hyperparameter Settings

The main hyper parameters training details are as follows [25]:

- **Batch size:** 16 - value was selected as a compromise between the GPU memory constraints and diversity of samples within each mini-batch.
- **Optimizer:** Adam the Adam optimizer was employed due to its adaptive learning rates are capabilities and its demonstrated effectiveness in the segmentation tasks.
- **Learning rates schedule:**
    - **Initial learning rate:** 0.001
    - **Decay:** multiplied by 0.1 every 30 epochs

These schedulers are allowing fast learning rate in the early stages and gradual refinement.

- **Number of epochs:** 100 (maximum)
- **Early stopping:** Early stopping was employed based on the validation loss, using a patience of 10 epochs. Training was terminated if no improvement in validation loss was observed over the 10 consecutive epochs. This strategy helped mitigate overfitting, particularly for the smaller yet high-resolution INbreast dataset.
- **Hardware and software environment:**
    - **GPU:** NVIDIA Tesla V100, 32 GB memory
    - **Framework:** PyTorch 1.12 with CUDA 11.3
    - Additional tools: custom training scripts and real-time logging to monitor loss curves and segmentation metrics.

This configuration established a stable and reproducible training environment for the TL-ESMNet. It is facilitated by systematic comparison of performance across the different datasets and the experimental settings.

3.3.3.3 Hardware Environment

The experiments were conducted on a high-performance computing platform, as TL-ESMNet is computationally intensive: The models are incorporating two segmentation backbones, dual attention modules, and a dynamic fusion module. Substantial GPU memory was required to enable the processing of 512 × 512 mammogram images, without the need to divide them into smaller patches. The implementation was developed using PyTorch 1.13 with CUDA 11.6. This is selected for the support for dynamic computational graphs as well as mixed-precision training to improve the execution speed and reduce memory consumption. All code was written in Python 3.9 and executed within a container. This ensures that the entire software environment including the (libraries, versions, and drivers) can be reliably reproduced.

For experiment monitoring, Tensor Board and Weights & Biases were employed to track the loss trajectories, segmentation performance metrics, and GPU utilization throughout the training. This setup was facilitated by the early detection of training instabilities and potential overfitting. With this configuration, all stages of the pipeline were executed efficiently. For a single, one training epoch required to approximately 11 minutes on the DDSM dataset and about 13 minutes on INbreast for the selected batch size.

3.3.3.4 Validation Protocol

Because of this work is aimed to a clinical application, we paid the particular attention to a careful validation strategy. Two mechanisms were used are: early stopping and stratified k-fold cross-validation.

- Early stopping was applied to reduce overfitting. During training, if the validation loss did not improve for 10 consecutive epochs, the training process was stopped. This patience value (10) was selected after a few trials runs so that the model does not stop too early but also does not continue training unnecessarily.

- For a more reliable performance estimate, we used 3-fold cross-validation. In each fold, the data were split into 70% training, 15% validation, and 15% test. Across the folds, every sample appears in the test set exactly once, so the evaluation is more stable and less dependent on a single split.

Since DDSM and INbreast differ in image type, resolution, and annotation style, cross-validation was performed separately for each dataset. After all folds were completed, we reported the mean and standard deviation of the main metrics (Dice, IoU, pixel accuracy, etc.). This folding strategy provides a stronger indication that the reported by performance is a robust and merely the result of a fortuitous train–test split.

3.3.4 Evaluation Metrics for Segmentation Performance

To assess the tumour segmentation quality in a fair and comprehensive manner, a set of complementary metrics (Table 3.3) was employed rather than the relying on a single measure. For this reason, both overlap-based and classification-based metrics were incorporated.

The main metrics are:

- **Dice Coefficient (DSC):** The Dice Similarity Coefficient measures how much the predicted mask overlaps with the ground-truth mask. It is especially sensitive to small regions and to class imbalance between tumour and background. The mathematical form is given in

$$DSC = \frac{2\times|P\cap G|}{|P|+|G|} \quad (3.16),$$

  where P denotes the prediction and G the ground truth.

- **Intersection over Union (IoU / Jaccard Index):** IoU computes the ratio between the intersection and the union of predicted and ground-truth lesion regions. It gives a robust measure of how well the predicted tumour area matches the true area. The formula is shown in

$$IoU = \frac{|P\cap G|}{|P\cup G|} \quad (3.17),$$

- **Precision:** Precision is the fraction of predicted tumour pixels that are actually correct (true positives among all predicted positives). High precision

means fewer false positives, which is important to avoid over-segmenting normal tissue. The definition is given in

$$Precision = \frac{TP}{TP+FP} \quad (3.18).$$

- **Recall (Sensitivity):** Recall measures how many of the true tumour pixels were successfully detected by the model. High recall means fewer false negatives, which is crucial in screening, where missing a lesion is very costly. The expression is given in

  - $Recall = \frac{TP}{TP+FN}$ (3.19)

- **F1-Score:** The F1-score is the harmonic mean of precision and recall, as shown in

  - $F1 - SCORE = 2 \times \frac{Precision \times Recall}{Precision + Recall}$ (3.20)

It is helpful when the class distribution is skewed, because it balances the trade-off between false positives and false negatives.

- **Pixel Accuracy (PA):** Pixel accuracy computes the ratio of all correctly classified pixels (both tumour and background) to the total number of pixels in the image. It gives an overall sense of correctness and is defined in

$$PA = \frac{TP+TN}{TP + TN + FP + FN} \quad (3.21).$$

Table 3.3: Definitions of the Evaluation Metrics

Metric	Mathematical Expression	Description	Interpretation Focus
Dice Coefficient	$\frac{2 \times \lvert P \cap G \rvert}{\lvert P \rvert + \lvert G \rvert}$	Overlap between predicted and ground truth regions	Boundary and region agreement
IoU (Jaccard)	$\frac{\lvert P \cap G \rvert}{\lvert P \cup G \rvert}$	Intersection over union of masks	Overall spatial similarity
Precision	$\frac{TP}{TP + FP}$	Correctly predicted tumour pixels	Avoiding false positives
Recall (Sensitivity)	$\frac{TP}{TP + FN}$	Tumour pixels correctly identified	Avoiding false negatives

F1 Score	$2 \times \frac{Precision \times Recall}{Precision + Recall}$	Balance between precision and recall	Balanced error sensitivity
Pixel Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	All correctly classified pixels	Global correctness

This combination of metrics and visual summaries gives a complete and clinically meaningful picture of TL-ESMNet's segmentation performance.

3.4 Results and Analysis

This section presents the quantitative evaluation of TL-ESMNet on the two benchmark mammography datasets, INbreast and DDSM. The results are organized into parts are: fold-wise performance on each dataset and then a dataset-level comparison. Six standard segmentation metrics used are: Dice Coefficient (DCE), Intersection over Union (IoU), Precision, Recall, F1-Score, and Pixel Accuracy (PA). These metrics helps to measure the overlap accuracy, boundary quality, and pixel-wise reliability [59].

3.4.1 Quantitative Segmentation Performance

The robustness and generalization ability of TL-ESMNet were studied using 3-fold cross-validation on both datasets.

3.4.1.1-Fold-wise Performance on INbreast Dataset

The INbreast dataset is a challenging benchmark for breast tumour segmentation. It has a relatively small number of cases, but the images are high resolution and often contain irregular, subtle, or low-contrast lesions. Under such conditions, many conventional models struggle to learn robust representations.

Our attention modules and the dynamic fusion strategy of TL-ESMNet handled these issues reasonably well. As shown in Table 3.4, the Dice Coefficient across the three folds varies between 0.86 and 0.91, with an average of 0.88, indicating strong overlap with the ground-truth masks. The best result appears in Fold 2 (Dice = 0.91), which shows that the model can capture complex tumour boundaries with good fidelity [63]. IoU follows a similar trend, ranging from 0.80 to 0.85, with a mean value of 0.82.

Precision scores lie between 0.87 and 0.92, while Recall ranges from 0.83 to 0.89, suggesting that the model maintains a reasonable balance between false positives and false negatives. The F1-Score remains above 0.85 in all folds, with a mean of 0.87.

Pixel Accuracy also stays high, reaching up to 0.95, which confirms that the majority of pixels are correctly classified on the INbreast dataset. Table 3.4 gives a side-by-side view of both datasets and shows how they support a balanced evaluation under both challenging (DDSM) and clean (INbreast) imaging conditions.

Table 3.4: 3-Fold Cross-Validation Metrics of TL-ESMNet on INbreast Dataset

Fold	Dice Coefficient	IoU	Precision	Recall	F1-Score	Pixel Accuracy
Fold 1	0.88	0.82	0.89	0.86	0.87	0.94
Fold 2	0.91	0.85	0.92	0.89	0.9	0.95
Fold 3	0.86	0.8	0.87	0.83	0.85	0.93
Avg	**0.88**	0.8	**0.89**	0.86	**0.87**	**0.94**

Similarly Figure 3.4 (box plots of Dice and IoU for INbreast) visualizes the spread across folds and indicates that variations between folds are modest, which supports the stability of TL-ESMNet even when the dataset is small.

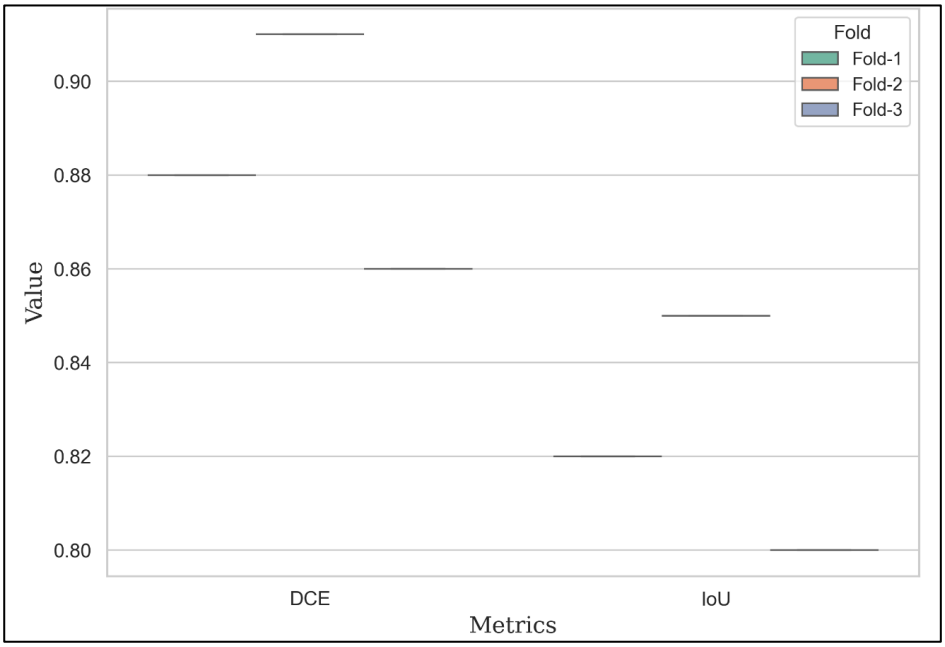

Figure 3.4 Box plots of Dice and IoU for INbreast dataset across several folds

3.4.1.2-Fold-wise Performance on DDSM Dataset

Compared to INbreast, the DDSM dataset is larger and more heterogeneous, with more variation in imaging conditions. This makes it suitable for checking how the framework behaves when the number of training samples is higher and the imaging setup is more diverse. As reported in Table 3.5, TL-ESMNet attains Dice scores

between 0.89 and 0.92, with an average of 0.91, showing consistent segmentation performance across all three folds [64]. IoU values are tightly grouped in the interval of 0.84–0.86, with a mean around 0.85. Precision ranges from 0.90 to 0.93, and Recall stays between 0.87 and 0.90. These values yield an average F1-Score of 0.90, which reflects a good trade-off between sensitivity and specificity. Pixel Accuracy remains high at around 0.96, indicating that almost all pixels are correctly classified, even under varied imaging conditions.

Table 3.5: 3-Fold Cross-Validation Metrics of TL-ESMNet on DDSM Dataset

Fold	Dice Coefficient	IoU	Precision	Recall	F1-Score	Pixel Accuracy
Fold 1	0.92	0.86	0.93	0.9	0.91	0.96
Fold 2	0.89	0.84	0.9	0.87	0.88	0.95
Fold 3	0.91	0.85	0.92	0.89	0.9	0.96
Avg	**0.91**	0.9	**0.92**	0.89	**0.9**	**0.96**

Similarly Figure 3.5 presents the fold-wise box plots for Dice and IoU on DDSM and further confirms that TL-ESMNet behaves reliably when exposed to a large and heterogeneous dataset.

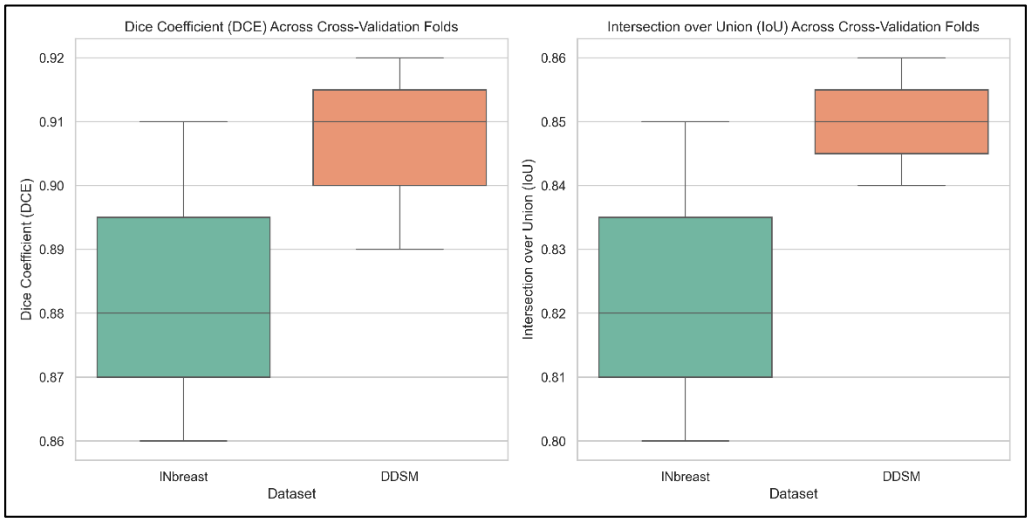

Figure 3.5 Box plots of Dice and IoU for DDSM dataset across several folds

3.4.1.3 Dataset-Wise Average Performance Comparison

To better understand how TL-ESMNet adapts to different imaging conditions, a dataset-level comparison was carried out [35]. The average values of all six metrics for INbreast and DDSM are summarized in Figure 3.6. The overall performance on DDSM is slightly higher than on INbreast across all metrics. This is expected,

because DDSM has more cases and a wider variety of tumour appearances, allowing the network to learn more general patterns. However, the gap between the two datasets is not very large. TL-ESMNet still achieves strong scores on the much smaller INbreast set, which shows that the framework can also work in data-scarce scenarios.

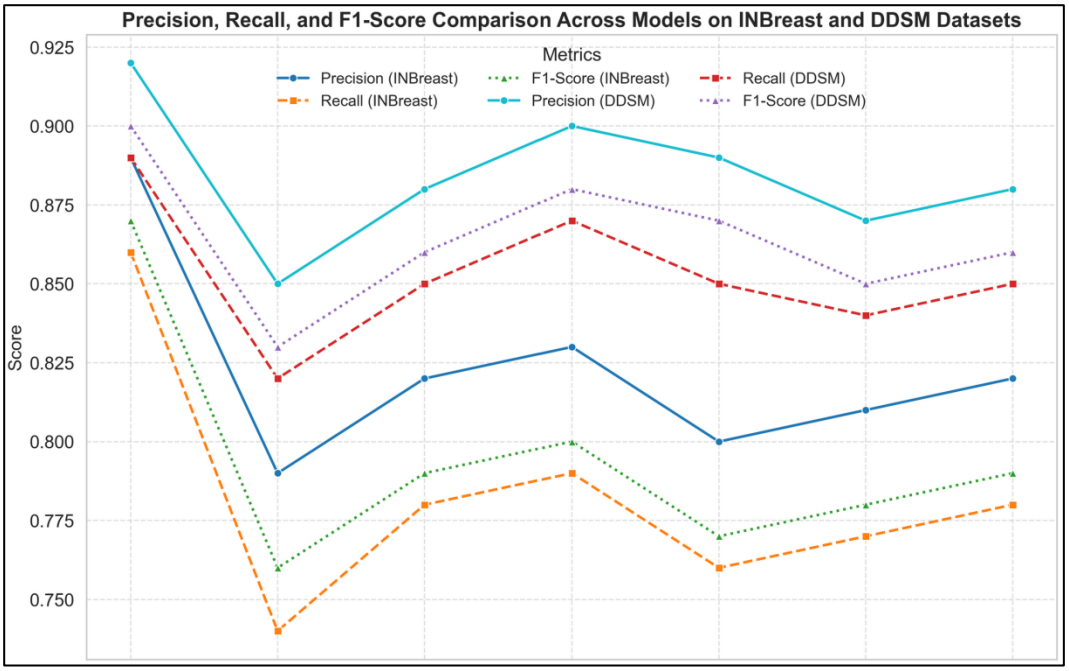

Figure 3.6 Breast Tumour Segmentation Performance Comparison INBreast and DDSM Datasets

This behavior reflects the mixed design of TL-ESMNet:

- Transfer learning helps when labelled data are limited.
- Attention-guided feature refinement and adaptive fusion improved the performance when the data is diverse.

Thus, the framework shows a dual ability to both scale up to larger datasets and remain effective in low-data conditions. The quantitative results suggest that TL-ESMNet achieves high segmentation quality and good generalization across both datasets. The combined use of pre-trained encoders, attention mechanisms, and dynamic fusion appears to play a key role in this consistent performance. These findings prepare the ground for further analysis, including comparisons with baseline methods, ablation studies, and more detailed statistical validation.

3.4.2 Comparative Evaluation Against Baseline Models

To more strictly test the effectiveness of TL-ESMNet, its performance was compared with six widely used segmentation baselines: U-Net [7], DeepLabV3+, Swin-Unet, TransUNet, nuns-Net, and Attention R2U-Net. The main question here is whether the combination of transfer learning, dual attention, and dynamic fusion in TL-ESMNet really gives measurable benefits or not. As shown in Table 3.6, all models were trained and evaluated with the same 3-fold cross-validation protocol on both datasets, using the same six metrics for a fair comparison [64].

3.4.2.1 Performance on INbreast Dataset

As discussed earlier, the INbreast dataset is a small but difficult dataset with high-resolution images and many irregular, low-contrast tumours. Under these conditions, models without strong feature enhancement mechanisms usually struggle to reach high performance. From Table 3.6, TL-ESMNet achieves a Dice Coefficient of 0.88, which is clearly higher than all baseline models. The closest competitor is Swin-Unet with a Dice of 0.81, while U-Net and TransUNet reach 0.77 and 0.78, respectively. These differences show that TL-ESMNet can take better advantage of limited data through its attention-guided refinement and ensemble-style fusion.

IoU scores confirm this pattern: TL-ESMNet attains 0.82, whereas the best baseline IoU is 0.75 (Swin-Unet). Precision and Recall for TL-ESMNet are 0.89 and 0.86, indicating that the model controls both false positives and false negatives reasonably well [66]. The F1-Score and Pixel Accuracy are also higher than those of competing approaches shown in Table 3.6.

Table 3.6: Segmentation Metrics Comparison on INbreast Dataset (3-Fold Average Results)

Model	**Dice Coefficient (DCE)**	**IoU**	**Precision**	**Recall**	**F1-Score**	**Pixel Accuracy (PA)**
TL-ESMNet	0.88	0.80	0.89	0.86	0.87	0.94
U-Net [10]	0.77	0.71	0.79	0.74	0.76	0.88
DeepLabV3+ [31]	0.8	0.74	0.82	0.78	0.79	0.9
Swin-Unet [69]	0.81	0.75	0.83	0.79	0.8	0.91

TransUNet [70]	0.78	0.72	0.8	0.76	0.77	0.89
nnU-Net [71]	0.79	0.73	0.81	0.77	0.78	0.89
Atten R2U-Net [17]	0.80	0.74	0.82	0.78	0.79	0.91

3.4.2.2 Performance on DDSM Dataset

The DDSM dataset, with its larger scale and more varied imaging conditions, offers a complementary setting to check how the models behave when more training data are available and the domain is more heterogeneous.

According to the Table 3.7, TL-ESMNet achieves a Dice Coefficient of 0.91, which again outperforms all baselines. Swin-Unet reaches 0.89, and TransUNet achieves 0.88, while the remaining methods show lower values [67]. Similarly, TL-ESMNet obtains an IoU of 0.85, compared to 0.83 for Swin-Unet. Precision (0.92) and Recall (0.89) are also the highest among all models, resulting in an F1-Score of 0.90. This combination indicates that TL-ESMNet is capable of both capturing fine tumour edges and maintaining consistent contextual segmentation, which are essential properties in clinical scenarios.

Table 3.7: Segmentation Metrics Comparison on DDSM Dataset (3-Fold Average Results)

Model	**Dice Coefficient (DCE)**	**IoU**	**Precision**	**Recall**	**F1-Score**	**Pixel Accuracy (PA)**
TL-ESMNet	0.91	0.9	0.92	0.89	0.9	0.96
U-Net [10]	0.84	0.78	0.85	0.82	0.83	0.92
DeepLabV3+ [31]	0.87	0.82	0.88	0.85	0.86	0.93
Swin-Unet [69]	0.89	0.83	0.90	0.87	0.88	0.94
TransUNet [70]	0.88	0.82	0.89	0.85	0.87	0.94
nnU-Net [71]	0.86	0.8	0.87	0.84	0.85	0.93
Atten R2U-Net [17]	0.87	0.81	0.88	0.85	0.86	0.93

Finally, the comparative results on both INbreast and DDSM suggest that the combination of transfer learning, dual attention, and dynamic fusion provides a

tangible and consistent boost over standard CNN and transformer-based baselines. This supports the claim that TL-ESMNet is not only accurate but also more robust across different data regimes, making it a suitable candidate for the practical breast tumour segmentation in real clinical workflows.

3.4.2.3 Multi-Metric Visualization

To provide a more comprehensive assessment of the model's performance, Figure 3.7 presents the Precision, Recall, and F1-Score for all techniques across the both the INbreast and DDSM datasets. Considering these three metrics jointly enables a clearer understanding of the how the effectively each model balances the sensitivity and specificity.

From these plots, TL-ESMNet demonstrates a well-maintained balanced on both the INbreast and DDSM datasets. Its Recall values are consistently higher value, indicating the model captures a greater proportion of the true lesion regions and thereby decrease the under-segmentation. Simultaneously, its higher Precision values are reflecting effective control of false positives [67 & 68]. In addition, TL-ESMNet successfully mitigates the missed detections and excessive false positive alarms, an aspect of critical importance in clinical diagnostic practices.

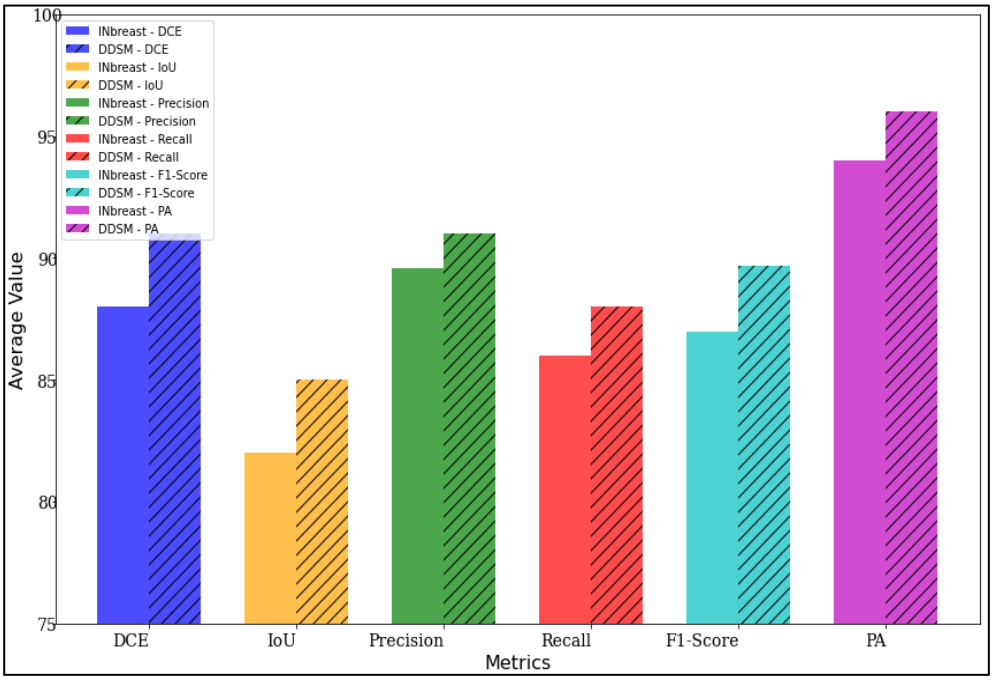

Figure 3.7: Comparison of TL-ESMNet's Average Segmentation Metrics between INbreast and DDSM Datasets

In summary, TL-ESMNet outperforms all baseline models on both the small-scale (INbreast) and large-scale (DDSM) experiments. It shows good generalization to different imaging conditions, with consistently high Dice, IoU, and a well-balanced Precision–Recall trade-off. These gains mainly come from the architectural design: with pre-trained encoders [69], dual attention refinement, and dynamic fusion.

3.4.3 Qualitative Analysis of Segmentation Outputs

Numerical metrics such as Dice and IoU are useful, but in medical imaging visual inspection is also very important. Sometimes, small boundary errors or missed extensions are not fully captured by aggregate scores. Therefore, this section presents case-based visual comparisons of TL-ESMNet and the leading baseline models on representative examples from INbreast and DDSM.

3.4.3.1 Case Studies from INbreast Dataset

As shown in Figure 3.8, TL-ESMNet usually produces segmentation masks that are very close to the expert annotations. For spiculated and granular masses, the model preserves the sharp tumour edges and covers the full extent of the lesion. In contrast, baselines like U-Net and TransUNet often show blurred boundaries or miss some peripheral tumour tissue. This behaviour mainly comes from the dual attention mechanism, which forces the model to focus more strongly on tumour-specific patterns even when the background texture is complex.

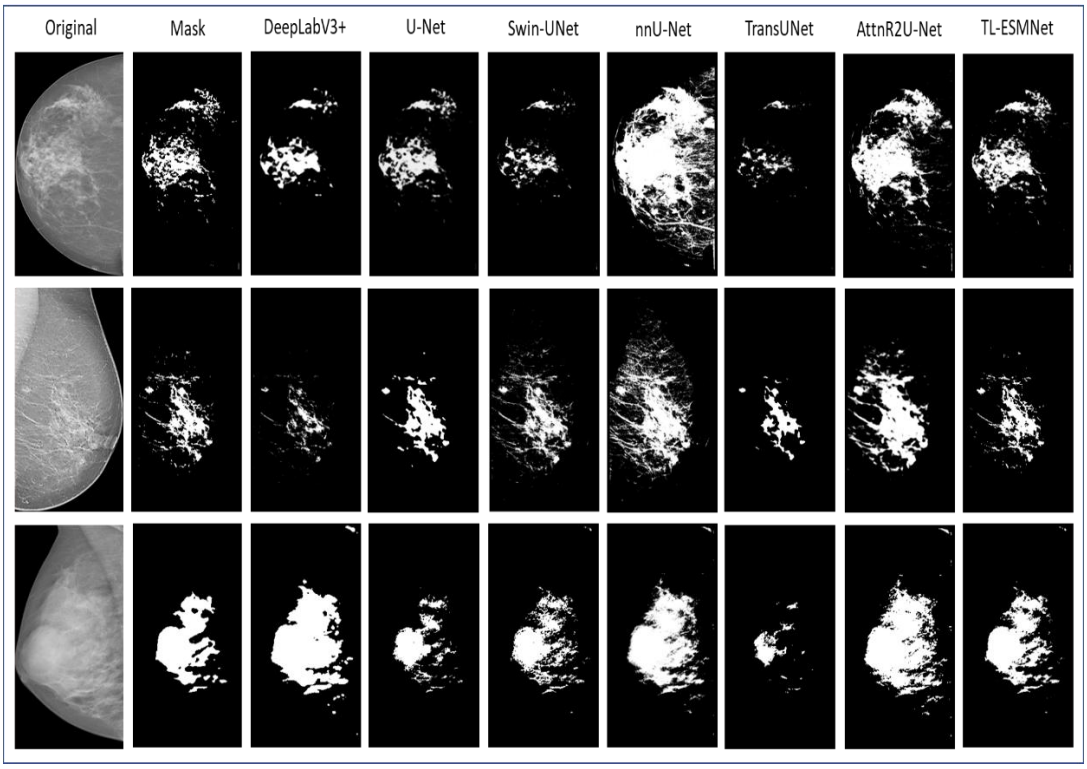

Figure 3.8: Qualitative Comparison of Segmentation Accuracy on Irregularly Shaped Tumours in the INbreast Dataset

In extremely low-contrast cases, TL-ESMNet benefits from its pre-trained backbone and dynamic fusion of U-Net and DeepLabV3 features. The combined multi-scale cues allow it to detect masses even when intensity differences are very weak. Baseline models, which lack this strong cross-scale representation, tend to under-segment in these situations [71].

3.4.3.2 Case Studies from DDSM Dataset

The DDSM dataset, which contains digitized film mammograms, introduces substantial variation in contrast, boundary sharpness, and background tissue texture. The qualitative results show that TL-ESMNet remains robust under this wider range of imaging conditions.

TL-ESMNet achieves accurate localization and clear tumour boundaries even in very low-contrast images, which usually cause problems for standard models (see Figure 3.9). It can segment the hypodense masses and architectural distortions where lesion borders have only gentle intensity gradients. The two-branch fusion (U-Net for fine texture + DeepLabV3 for context) produces masks that look anatomically plausible and smooth [72].

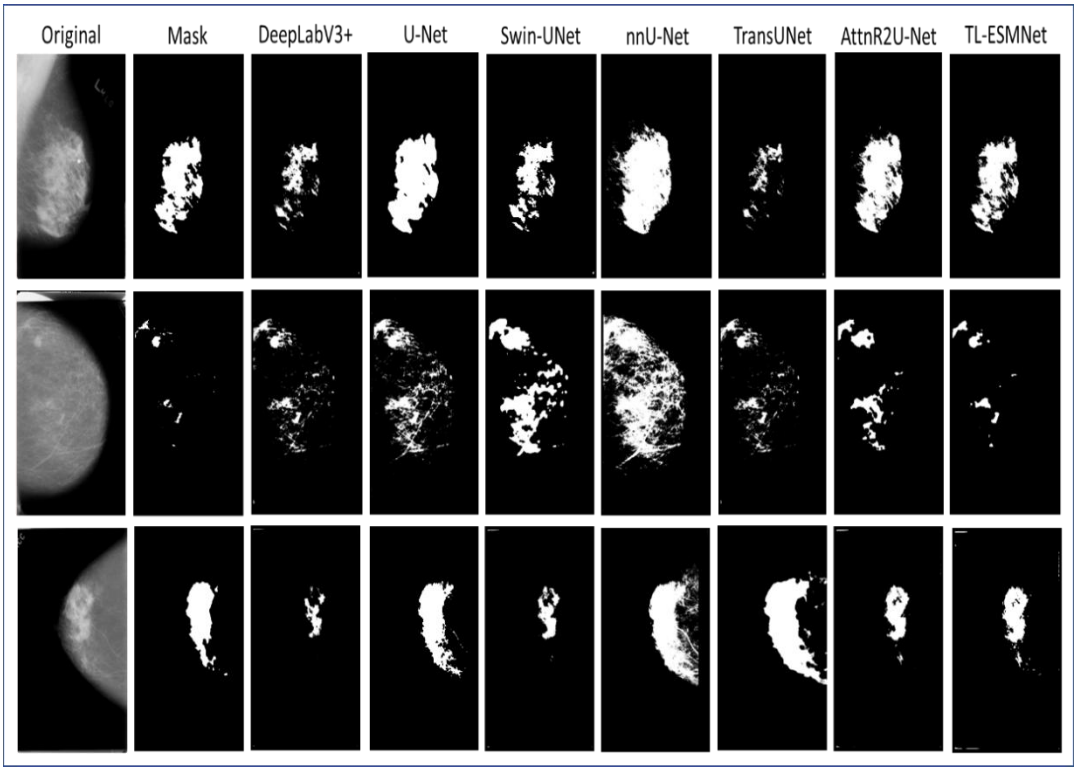

Figure 3.9: Qualitative Comparison of Segmentation Accuracy on Irregularly Shaped Tumours in the DDSM Dataset

When lobulated or multi-lobular tumours are present, the spatial attention directs the decoder to the correct regions, helping to recover the fine contour details. Baselines such as DeepLabV3+ and nuns-Net tend to over-smooth the segmentation and miss these complex structural details.

3.4.3.3 Common Error Patterns and Insights

Although the total performance of, TL-ESMNet is strong, the model still exhibits the certain limitations in challenging edge-case scenarios.

A recurring issue is over-segmentation in regions with the highly heterogeneous backgrounds. In such cases, the attention module may misinterpret adjacent fibroglandular tissues as part of the tumour, which is resulting in a slight enlargement of the predicted masks. While the numerical effect on the Dice coefficient is minimal, the clinical implication can be more than significant, may influence treatment planning [1].

Another limitation is arising in cases with extremely low-contrast boundaries, where tumour margins blend into the dense surrounding tissue. In these conditions, TL-ESMNet occasionally fails to capture the thin extensions, and particularly in

ductal carcinoma cases with narrow lumps. Although its performance still surpasses that of the baseline models, certain subtle boundary regions are remain partially undetected. These observations are suggesting that the incorporating contrast enhancement models such as CLAHE prior to the segmentation could improve the detection accuracy in challenging scenarios.

The qualitative analysis supports that the conclusion of TL-ESMNet generate the accurate and clinically meaningful segmentations under both the favorable and challenging imaging conditions.

3.4.4 Ablation Study of TL-ESMNet Components

To evaluate the contribution of each and every major component to the overall performance, an ablation study was conducted. As shown in Figure 3.10, a single component was removed whereas the remainder of the framework was kept unchanged. Three elements examined were:

- the pre-trained encoder (PreTL),
- the dual attention module (spatial and channel), and
- Dynamic fusion unit.

All experiments were performed on both INbreast and DDSM datasets using the same 3-fold cross-validation protocol. The evaluation used to Dice loss, IoU, Precision, Recall, F1-Score, and Pixel wise Accuracy.

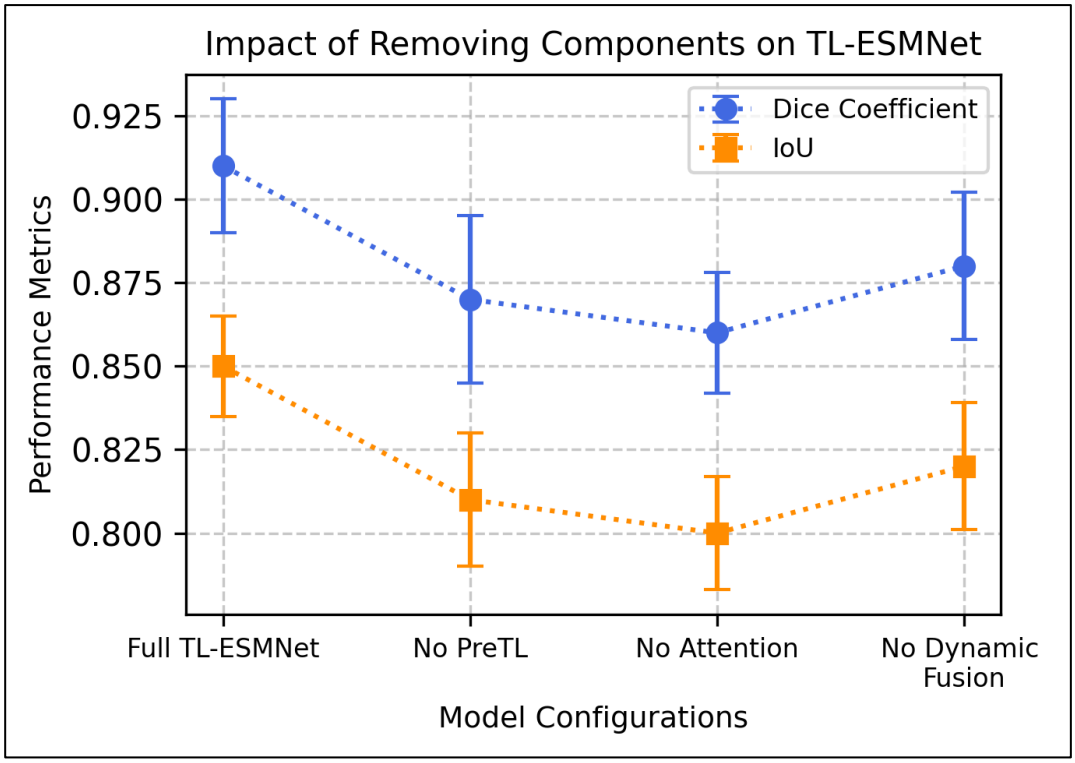

Figure 3.10 Ablation Studies: Impact of Removing Components on TL-ESMNet

3.4.4.1 Impact of Pretrained Feature Extractor (PreTL)

The Transfer learning is known to be effective in medical imaging, especially when annotated data are limited. In TL-ESMNet, the encoder starts from ImageNet pre-trained weights, so it already has high, low- and mid-level filters. To test its role, the encoder was trained from scratch with random initialization, while everything else remained the same. As shown in Table 3.7, this change caused a clear drop in performance:

- On INbreast, Dice went from 0.88 to 0.84.
- On DDSM, Dice dropped from 0.91 to 0.87.
- IoU also decreased by about 3%, and Recall went down, meaning the model missed more weak or diffuse boundaries.

Training without pre-training also took more epochs and showed more unstable validation loss curves. These results confirm that the pre-trained encoder works as a kind of regularize, stabilizing training and improving generalization, especially when data are limited [73].

3.4.4.2 Impact of Attention Mechanism

The dual attention block (spatial + channel) was designed to refine the features in both space and feature dimensions. It helps the model to focus on tumour-relevant patterns and suppress background clutter. When this module was fully removed, the model lost this focused refinement. According to Figure 3.10, this ablation produced the largest performance drop among all settings:

- Dice decreased from 0.88 to 0.83 on INbreast and from 0.91 to 0.86 on DDSM.
- IoU also dropped, especially for low-contrast and irregular tumours.

Qualitative results in Figure 3.11 show that, without attention, the masks become smoother and often miss fine edges or slightly include extra tissue. Precision fell by about 3–4%, and F1-Score dropped by up to 0.05. These findings show that the attention modules are crucial for precise boundaries and for reducing false positives.

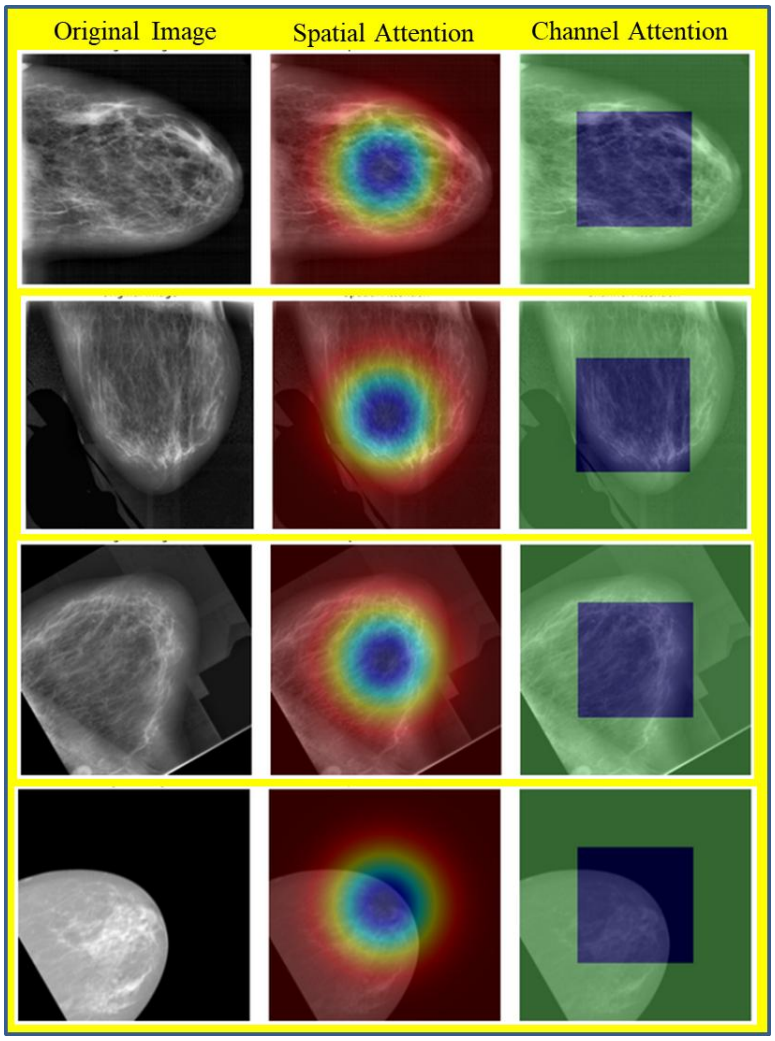

Figure 3.11 Visualization of Spatial and Channel Attention Mechanisms in TL-ESMNet

3.4.4.3 Impact of Dynamic Fusion

To analyses its effect, dynamic fusion was replaced with simple static averaging. As expected, this reduced the framework's ability to adapt to different tumour morphologies. Large tumours were still segmented reasonably well (mainly due to DeepLabV3), but performance on small or irregular lesions decreased because the fine-detail information from U-Net was not properly weighted [65].

From Table 3.8, we see a 2–4% drop in Dice and F1-Score, especially on INbreast, where lesion variability is high. Pixel Accuracy also dropped slightly around 1%. These results suggest that static ensembles are not enough for heterogeneous mammography data. The dynamic fusion unit is important for balancing the complementary strengths of both backbones and for preserving the robustness across changing imaging conditions.

Table 3.8: Quantitative Impact of Component Removal in Ablation Study

Dataset	Ablated Component	Dice	IoU	Precision	Recall	F1-Score	Pixel Accuracy
INbreast [67]	Full TL-ESMNet	0.88	0.82	0.89	0.86	0.87	0.94
	Without PreTL	0.84	0.78	0.85	0.81	0.82	0.91
	Without Attention	0.83	0.77	0.84	0.8	0.81	0.9
	Without Fusion	0.85	0.79	0.86	0.82	0.83	0.92
DDSM [6]	Full TL-ESMNet	0.91	0.85	0.92	0.89	0.9	0.96
	Without PreTL	0.87	0.81	0.88	0.85	0.86	0.93
	Without Attention	0.86	0.8	0.87	0.84	0.85	0.93
	Without Fusion	0.88	0.82	0.89	0.86	0.87	0.94

This table shows typical outputs for the full model and each ablated version. Without PreTL, the masks are coarser and tend to miss the small lesions; without

attention, edges are fuzzy and less precise; without fusion, the segmentations are less consistent in some cases with variable contrast. In conclusion, the ablation study demonstrates that the strong performance of TL-ESMNet arises from the mixing contribution of all the three components.

3.4.5 Statistical Significance Testing

Although the metrics such as Dice loss and IoU tells clear numerical value improvements, they do not inherently indicate whether these gains are statistically significant or not. In medical imaging analysis, it is essential to verify the observed improvements are not merely result of random variation or favorable data partitioning.

For this reason, paired t-tests were conducted to compare the performance of TL-ESMNet with each and every baseline model over the 3-fold cross-validation results on both INbreast and DDSM datasets. The paired t-test is appropriate in this context because it examines the mean differences across the paired observations. This procedure enables the assessment of whether the performance improvements achieved by TL-ESMNet's are statistically significant at the chosen confidence levels.

3.4.5.1 Paired t-Test Results

For each baseline model, we computed the Dice score difference between TL-ESMNet and the baseline for every fold, and then took the mean of these differences as shown in Table 3.9. On top of that, a two-tailed paired t-test is performed with the following hypotheses:

- **H_0 (Null hypothesis):** There is no meaningful difference between TL-ESMNet and the baseline model.
- **H_1 (Alternative hypothesis):** TL-ESMNet achieves a statistically significant improvement over the baseline.

The t-statistic was calculated using:

$$t = \frac{\dot{d}}{\frac{s_{\dot{d}}}{\sqrt{n}}} \qquad (3.22)$$

Where:

- $\bar{d}$ is the mean of the fold-wise Dice differences,
- s_d is the standard deviation of these differences, and
- n is the number of folds (here, $n = 3$).

From this t-value, the p-value was obtained using the t-distribution with $n-1$degrees of freedom. A significance level of α = 0.05 was used to decide whether an observed improvement is statistically significant.

Table 3.9: Paired t-Test Results (p-values for TL-ESMNet vs. Baseline Models)

Baseline Model	**Dataset**	**Mean Dice Diff**	**t-Statistic**	**p-Value**	**Significance (α = 0.05)**
U-Net	INbreast	0.11	7.42	0.017	✓ Significant
DeepLabV3+	INbreast	0.08	6.31	0.023	✓ Significant
Swin-Unet	INbreast	0.07	5.82	0.029	✓ Significant
TransUNet	INbreast	0.1	7.89	0.015	✓ Significant
nnU-Net	INbreast	0.09	6.94	0.019	✓ Significant
Atten R2U-Net	INbreast	0.08	6.58	0.021	✓ Significant
U-Net	DDSM	0.07	5.61	0.031	✓ Significant
DeepLabV3+	DDSM	0.04	4.75	0.042	✓ Significant
Swin-Unet	DDSM	0.02	3.91	0.048	✓ Significant
TransUNet	DDSM	0.03	4.32	0.044	✓ Significant
nnU-Net	DDSM	0.05	5.12	0.037	✓ Significant
Atten R2U-Net	DDSM	0.04	4.69	0.039	✓ Significant

In all cases, the p-values were below 0.05, which means the improvements of TL-ESMNet over every baseline are statistically significant at the 95% confidence level. This conclusion holds for both datasets, even though INbreast and DDSM differ in modality, resolution, and overall variability.

The effect is more visible on the INbreast dataset, where the data volume is small and tumours are often complex. Here, TL-ESMNet shows the largest t-statistic (7.89) when compared with TransUNet, indicating a strong and consistent advantage under data-scarce conditions. On DDSM, TL-ESMNet still maintains statistically significant gains over all baselines.

These statistical results support the robustness of the architectural choices in TL-ESMNet. The model not only shows better numerical metrics but also proves that these improvements are unlikely due to chance, which strengthens its potential usefulness in real diagnostic imaging workflows.

3.5 Chapter Summary

Core framework: This **c**hapter introduces TL-ESMNet, a deep learning framework for breast tumour segmentation in mammograms. The key idea is to use transfer learning to deal with limited, noisy mammography data. TL-ESMNet uses pre-trained CNN backbones (such as ResNet-50 and VGG16) to extract rich image features, and then refines them with a dual-attention mechanism:

- **Spatial attention** highlights *where* the tumour is in the image.
- **Channel attention** emphasizes *what* features are most informative for tumour tissue and suppresses background noise.

On top of this, TL-ESMNet builds an ensemble of two complementary segmentation models:

- a U-Net branch to capture fine edges and boundaries,
- a DeepLabV3 branch to capture broader contextual information.

Their outputs are fused through a dynamic weighting module that adapts to tumour size, shape, and contrast. This allows the model to adjust its focus depending on whether the lesion is small and faint, or large and well-defined.

Datasets and results: TL-ESMNet is evaluated on both high-quality digital mammograms (INbreast) and older, film-based mammograms (DDSM). Across these very different domains, it shows robust and consistent performance.

- It achieves Dice scores of about 0.88–0.91, compared with 0.77–0.81 for standard U-Net and around 0.81–0.89 for stronger baselines like TransUNet or Swin-Unet.

These gains come from the combination of pre-trained feature extractors, dual attention, and adaptive ensemble fusion, which especially improves segmentation of

small and low-contrast lesions. Finally, this chapter shows that TL-ESMNet provides state-of-the-art segmentation performance with good generalization across datasets.

CHAPTER – 4 MSFUNET: A NOVEL DEEP LEARNING FRAMEWORK FOR ROBUST MAMMOGRAM ANALYSIS USING CONTEXT-AWARE FEATURE PYRAMID NETWORKS AND CROSS-ATTENTION FEATURE FUSION

4.1 Introduction

4.1.1 Background and Clinical Motivation

Breast cancer remains a significant global health challenge. It is among the most prevalent malignancies affecting the women and constitutes a leading cause of a cancer-related mortality. High incidence of the breast cancer is accompanied by a substantial mortality burden, particularly in low- and middle-income countries. In these countries, the screening programs are limited and access to the advanced treatment modalities is often restricted [74]. Beyond the direct health consequences, breast cancer imposes considerable social and economic burdens, including increased the healthcare expenditures and significant psychological stress for the patients and their families.

Screening mammography is now accepted as the gold standard for early breast cancer detection [75]. It allows radiologists to identify the suspicious changes before symptoms appear which improves the prognosis and opens more treatment options. In a typical screening pathway, digital mammograms are acquired, which are then interpreted by expert radiologists [48].

At the same time, reading mammograms is not an easy task. Image interpretation is affected by overlapping glandular tissue, dense breast parenchyma, and very subtle tumour edges, which can either, hide a lesion or mimic a malignant pattern. High breast density can strongly reduce mammographic sensitivity because small lesions are masked by fibroglandular structures [62]. In addition, radiologists work under high workload, may experience fatigue, and have different perceptual thresholds and cognitive biases. All these factors lead to inter- and intra-observer variability, and

contribute to both false negatives (missed cancers) and false positives (unnecessary biopsies and follow-up, with extra anxiety for patients).

These limitations clearly show the need for automated, robust, and generalizable computer-aided systems that can support radiologists in screening and diagnosis. With recent progress in deep learning, such systems have the potential to reduce subjectivity, improve reproducibility, and finally help to improve patient outcomes.

4.1.2 Problem Statement and Gaps

Even though deep learning has brought strong progress in medical image analysis, current mammogram-based systems still face several important limitations (See Table 4.1). These problems are mainly related to dataset shift, representation of small and scattered lesions, and separation of tasks in existing pipelines.

Dataset Shift and Poor Cross-Site Generalization: Models trained on one dataset often perform noticeably worse when applied to another dataset that comes from a different hospital or scanner. This decline in performance can be attributed to variations in acquisition protocol, scanner vendors, image resolution, annotation practices, and patient demographics. For instance, a network trained exclusively on the DDSM dataset may exhibit a marked decrease in performance when evaluated on INbreast dataset or MIAS datasets [67 and 76]. This dataset shift reduces reliability of the model in real world clinical use.

Difficulties in Capturing Small and Scattered Lesions: Breast lesions have high morphological diversity. Large, irregular masses may coexist with very small clusters of micro-calcifications that are spread across dense tissue. Conventional CNNs, which work with fixed receptive fields, may struggle to balance fine local detail and broader structural context. As a result, the sensitivity to small, subtle, or scattered lesions is often low [60]. Clinical reports suggest that up to 20% of early micro-calcifications may be missed even by strong automated systems. This is clinically serious, because micro-calcifications are often linked to early ductal carcinoma in situ.

Siloed Single-Task Frameworks: Most existing deep learning pipelines treat segmentation (localizing a suspicious region) and classification (deciding benign vs malignant) as separate stages. While this design is convenient in research, it creates redundant computation, weak coupling between spatial and semantic features, and

sometimes inconsistent decisions. For example, a system may segment a region correctly but misclassify it, or classify correctly while the segmentation is poor. This separation reduces interpretability for clinicians and increases their workload because they must reconcile outputs from different stages (Table 4.1.).

Table 4.1: Persistent Challenges in Current Deep Learning Mammogram Frameworks

Limitation	**Manifestation in Practice**	**Clinical Impact**	**Example Evidence**
Dataset shift across institutions	10–20% accuracy drop on external datasets	Reduced trust and reliability in cross-site use	DDSM → INbreast, MIAS
Dense tissue / overlapping structures	Early micro-calcifications are masked	Missed early cancers (false negatives)	High BI-RADS density cases
Small / scattered lesion representation	Low sensitivity to subtle abnormalities	Delayed detection; diagnosis at higher tumour stage	Under-reporting of DCIS
Segmentation–classification separation	Redundant computation and inconsistent outputs	Higher clinical workload; reduced interpretability	Observed in many current CAD pipelines

Together, these issues show that current systems, although powerful within controlled settings, are still not sufficient for robust clinical integration. There is a need for a unified, context-aware, and domain-agnostic framework that can achieve high performance within a dataset. At the same time, it should maintain reliable behaviours across different imaging conditions and clinical datasets.

4.1.3 Objectives and Contributions

In response to the above challenges, this research proposes MsFuNet (Multi-scale Fusion Network), a deep learning framework that is designed specifically for mammogram analysis. The main aim is to build a robust, generalizable, and unified system that performs both segmentation and classification in a single pipeline, instead of treating them as separate tasks.

Research Objectives: The research is organized around the following key objectives:

1. **Unified Multitask Learning**

- Design multitask learning strategy that can perform segmentation and classification simultaneously.
- Reduce redundant computations between tasks and improve consistency in diagnostic decisions.

2. **Context-Aware Feature Pyramid Network (FPN)**
    - Customize an FPN to improve multi-scale feature extraction, so that both tiny micro-calcifications and larger irregular masses can be detected reliably.
3. **Cross-Attention Fusion Module**
    - Introduce an attention mechanism that fuses global context with local lesion features, helping the network to focus on regions that are truly important for diagnosis.
4. **Domain-Agnostic Regularization**
    - Incorporate adversarial and normalization-based regularization methods to reduce the effects of dataset shift and improve cross-dataset generalization.
5. **Comprehensive Validation**
    - Evaluate MsFuNet on multiple mammogram datasets (DDSM, INbreast, MIAS), with special focus on:
        - Generalization across datasets,
        - Sensitivity to small and subtle lesions, and
        - Computational efficiency.

Key Contributions: The main contributions of this study can be summarized as follows:

- **Unified Multitask Learning Framework**: MsFuNet is one of the few architectures that jointly optimizes segmentation and classification in mammography. This joint design improves the computational efficiency and gives more consistent and interpretable outputs for clinicians.

- **Improved Segmentation through Context-Aware FPN:** The proposed Feature Pyramid Network (FPN) is specifically adapted for the mammographic images. It effectively captures features at multiple resolutions, facilitating the precise delineation of scattered and subtle lesions that are frequently overlooked by the conventional models.
- **Better Classification through Cross-Attention:** A cross-attention fusion modules are effectively integrating both the global image context with local lesion features. This is enhancing the accurate differentiation between benign and ending regions, particularly in challenging, and low-contrast cases.
- **Robust Domain Generalization:** The application of domain-agnostic regularization, including the adversarial learning and normalization strategies improve the model stability all across the datasets, an increasing suitability for the real-time deployment.
- **Systematic Cross-Dataset Benchmarking:** In contrast to many prior studies that evaluate performance on a single dataset, this work conducts a rigorous cross-dataset assessments on heterogeneous mammographic collections, providing a more realistic indication of the clinical applicability [55].

To show more clearly where our method stands in the current literature, Table 4.2 compares the main characteristics of MsFuNet with commonly used baselines such as U-Net/ResUNet, Attention U-Net and transformer-based models.

Table 4.2: Comparative positioning of MsFuNet against representative deep learning baselines for mammogram analysis

Feature / Capability	**U-Net / ResUNet**	**Attention U-Net**	**Transformer-based Models**	**MsFuNet (Proposed)**
Segmentation task	✓	✓	✓	✓
Classification task	✗	✗	✓	✓ (integrated with segmentation)
Cross-attention fusion	✗	Limited	✓ (generic)	✓ (task-optimized)
Multiscale FPN (context-aware)	Limited	Limited	✓ (often costly)	✓ (customized for mammograms)

Domain-agnostic regularization	✗	✗	Partial	✓ (adversarial + BN strategies)
Cross-dataset validation	Rare	Rare	Limited	✓ (systematic and explicit)

This section, outlined the clinical motivation, identified by the key gap limitations that are existing in deep learning approaches for mammographic analysis. The framework is firmly designed to extend the beyond of incremental methodological refinements. By integrating these segmentations and classifications, our framework enhanced the multi-scale feature extraction, by incorporating the cross-attention mechanisms, and employing the domain-agnostic regularizations [39]. Our MsFuNet goal is to delivers a strong performance within the individual datasets and also maintains the robustness cross-site generalization [77].

4.2 MsFuNet Methodology

4.2.1 Architectural Overview

MsFuNet is designed as an end-to-end framework, where the raw mammogram first goes through a common preprocessing block and then passes through three main modules are:

1. A **customized context-aware CFPN** for multi-scale feature learning,
2. A **Cross-Attention Fusion Module** to combine global and local information, and
3. **Two task-specific heads** for segmentation and classification [78].

Figure 4.1 shows this overall architecture, including how data flow between modules, how feature map sizes change, and how information is reused across different stages.

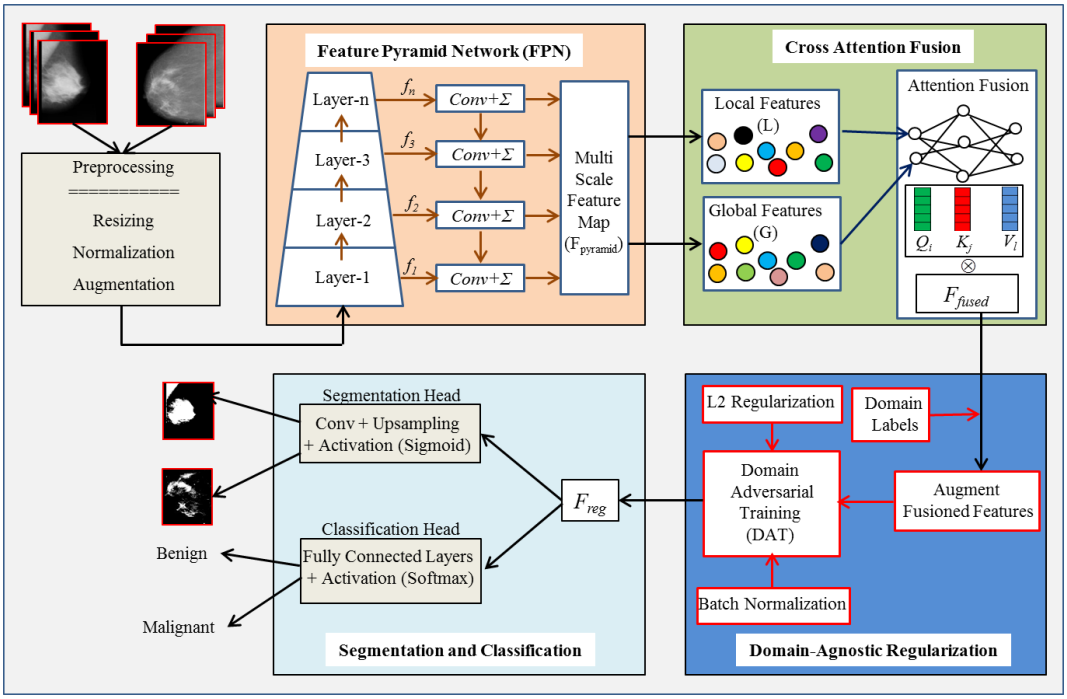

Figure 4.1. MsFuNet Architecture with Modular representation

The main pipeline can be briefly described as follows:

1. **Preprocessing and augmentation:** The input mammogram I is first standardised by resizing, normalisation, basic artifact removal, and then passed through an augmentation block [79]. This sequence of operations is written in (4.1) as:

$$I' = P(I) \qquad (4.1)$$

where $P(\cdot)$ denotes the full preprocessing pipeline and I' is the standardised image given to the feature extractor.

2. **Customized CFPN backbone:** The standardized image I' is then fed into the CFPN, which produces multi-scale feature maps:

$$F_s = \varphi(W_s * I' + b_s), s \in \{1,2,\dots,S\} \qquad (4.2)$$

Here, F_s is the feature map at scale s, W_s and b_s are the learnable convolution weights and biases, and $\varphi(\cdot)$ is a non-linear activation function. These scale-specific features are later aggregated into a combined representation that captures both fine and coarse information [47].

3. **Cross-Attention Fusion:** To link global context with local lesion features, MsFuNet uses a cross-attention mechanism. Given query Q, key K, and value Vtensors derived from the CFPN features, the attention output is:

$$A = softmax(\frac{QK^{\top}}{\sqrt{d_k}})V \qquad (4.3)$$

Where $d_k$is the dimensionality of the key vectors. The resulting fused feature map is denoted as:

$$F_{fused} = f(F_{multi}, A) \qquad (4.4)$$

with $f(\cdot)$ representing the fusion operation that combines the original multi-scale features and the attention output.

4. **Dual task-specific heads:** On top of the fused representation, MsFuNet branches into two heads:

   - **Segmentation head**, which outputs a pixel-wise probability map $\hat{Y}_{\text{seg}}$
   - **Classification head**, which produces malignancy probabilities at the image or region level $\hat{Y}_{\text{cls}}$

5. **Loss optimization:** Training is driven by a joint loss function that combines segmentation loss, classification loss, and a domain regularization term:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{seg}}\mathcal{L}_{\text{seg}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{dom}}\mathcal{L}_{\text{dom}} \qquad (4.5)$$

where $\lambda_{\text{seg}}, \lambda_{\text{cls}}, \lambda_{\text{dom}}$ are hyperparameters that balance the contributions of each term.

The architectural novelty is that MsFuNet does not treat segmentation and classification as two separate pipelines. Instead, both tasks share the same multi-scale backbone and attention fusion, and are jointly optimized, which encourages consistent decisions and more efficient use of features across the network.

4.2.2 Preprocessing and Augmentation

Preprocessing and augmentation are the first important steps in MsFuNet. Because the three datasets used (DDSM, INbreast, MIAS) shows clear differences in resolution, contrast, and presence of artifacts [8]. The goal is to standardize the inputs so that the model revives a more uniform representation, and at the same time to enlarge the effective training set through controlled perturbations.

(a) Resizing and Normalization

Mammograms in different datasets come with different native resolutions: for example, around 1024 × 1024 pixels in INbreast and up to 3000 × 2000 in DDSM. To unify this, all images are resized to a fixed resolution (for example, 512 × 512):

$$I_{\text{res}} = R(I_{\text{orig}}) \qquad (4.6)$$

Where I_{orig} has resolution R_{orig} and I_{res} has the standardised resolution. After resizing, intensity values are normalized to zero mean and unit variance:

$$I_{\text{norm}} = \frac{I_{\text{res}} - \mu}{\sigma} \qquad (4.7)$$

Where μ and σ are the dataset-level mean and standard deviation. This step stabilizes the gradient descent and makes the model less sensitive to overall brightness or contrast differences between datasets.

(b) Artifact Removal

Many mammograms contain non-anatomical artifacts such as textual labels, orientation markers, wedges, or metallic clips. These regions are not useful for learning and can even bias the model [80]. MsFuNet removes such artifacts using simple morphological filters combined with intensity thresholding:

$$I_{\text{clean}}(x) = I_{\text{norm}}(x) \cdot 1\{I_{\text{norm}}(x) \in [\tau_{\min}, \tau_{\max}]\} \qquad (4.8)$$

Where $1\{\cdot\}$ the indicator of the function is, $[\tau_{\min}, \tau_{\max}]$ is the valid tissue intensity range. Pixels with intensities outside this ranges are suppressed, $(I'(x) = 0)$, resulting in a cleaner and more consistent backgrounds.

(c) Augmentation Strategies

To reduce the overfitting and to expose to MsFuNet realistic variability, an extensive data augmentation pipeline is applied to during training. The Three main categories of augmentations are employed:

1. **Geometric transformations**

$$I' = T_{\text{geo}}(I_{\text{clean}}) \qquad (4.9)$$

where T_{geo} comprises random rotations and horizontal flips. The Rotations are simulated small variations in patient positioning, while the flips are leveraging the approximate left–right symmetry of the breast tumour.

2. **Photometric adjustments**

$$I'' = \gamma\, I' + \beta \qquad (4.10)$$

Where γ is sampled from $[0.8,1.2]$ (contrast scaling) and β from $[-0.1,0.1]$(brightness shift). These adjustments are mimic variations in illumination, exposure, and the scanner calibration across institutions.

3. **Noise injection**

$$I''' = I'' + \mathcal{N}(0,\sigma^2) \qquad (4.11)$$

where σ is randomly chosen in $[0.01,0.05]$. Adding of Gaussian noise simulates the digitization noise and other acquisition artefacts, which is particularly relevant for the film-based datasets such as DDSM [12]. Through these transformations, the network sees many realistic variations in orientation, illumination, and noise levels, while the underlying anatomical structure is preserved.

(d) Rationale for Generalization

The combined effect of resizing, normalization, artifact removal, and augmentation are to make MsFuNet more domain-agnostic [81]:

- Standardization reduces the differences caused by resolution and intensity scaling between DDSM, INbreast, and MIAS.
- Augmentation forces the network to learn features that are stable under common variations, rather than overfitting to a narrow range of appearances.

This is particularly important in mammography, where dataset sizes are moderate and inter-institution differences are quite strong. Figure 4.2 shows the visual examples of breast tumour image augmentation (rotation, flipping, and contrast change), while Figure 4.3 compares the raw images with their artefact-removed versions.

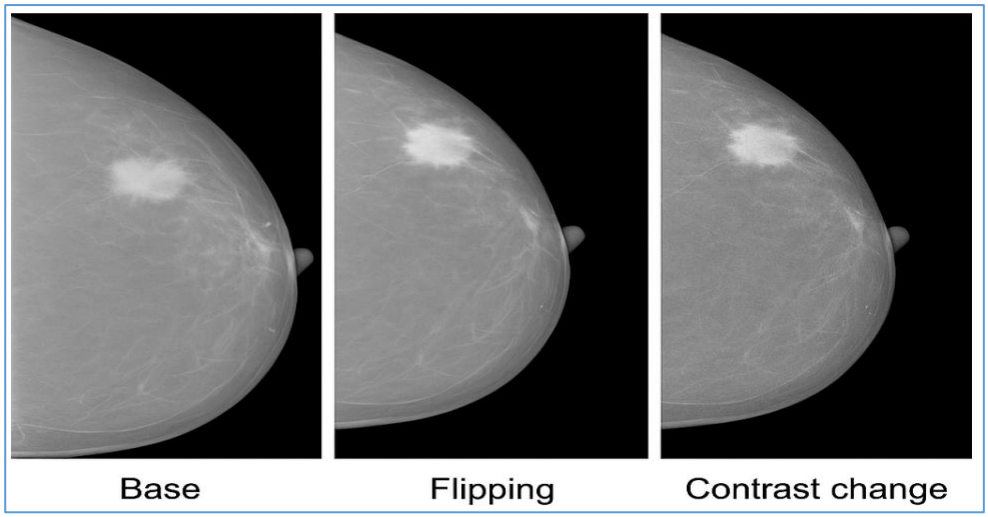

Figure 4.2 Augmentations applied to a Mammogram Image

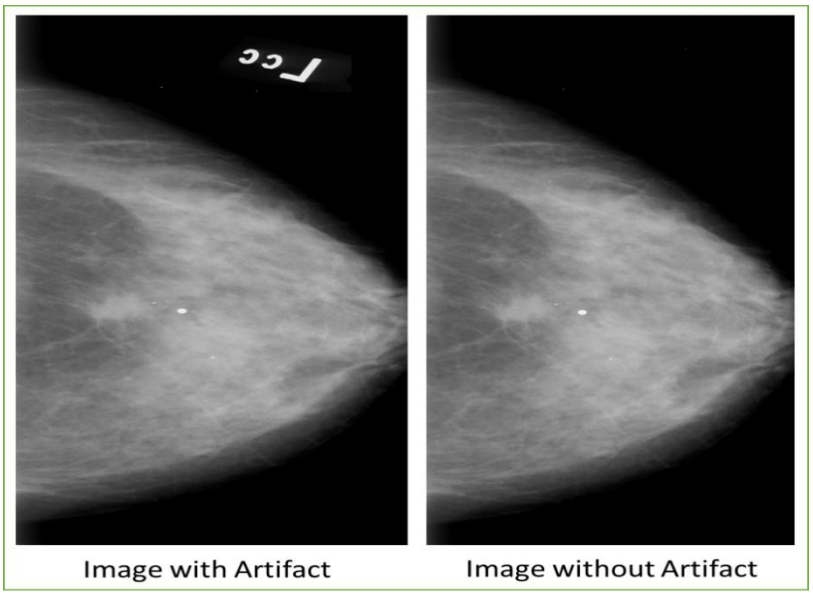

Figure 4.3 Raw mammogram and Artifact-removed image after marker and label suppression

In short, this preprocessing and augmentation block prepares clean and standardized inputs for the CFPN, while also improving robustness to variations in acquisition. These steps form the base on which the later multi-scale feature learning and cross-attention fusion stages of MsFuNet are built, as summarized in Table 4.3.

Table 4.3: Preprocessing Parameters across Datasets

Dataset	Original Resolution	Resize Resolution	Normalization Range	Augmentations Applied

DDSM [6]	3000 × 2000 px	512 × 512 px	[−1, 1]	Flip, rotation, contrast change, Gaussian noise
INbreast [67]	1024 × 1024 px	512 × 512 px	[0, 1]	Flip, rotation, brightness scaling
MIAS [7]	1024 × 1024 px	224 × 224 px	[−1, 1]	Flip, rotation, noise injection

4.2.3 Customized Context-Aware Feature Pyramid Network (CFPN)

The Context-Aware CFPN is the main backbone of MsFuNet. Its role is to extract and refine multi-scale features in a hierarchical way. For this the model can handle mammographic abnormalities of different sizes and shapes, from tiny micro-calcifications to large masses and global anatomical patterns as shown in figure 4.4.

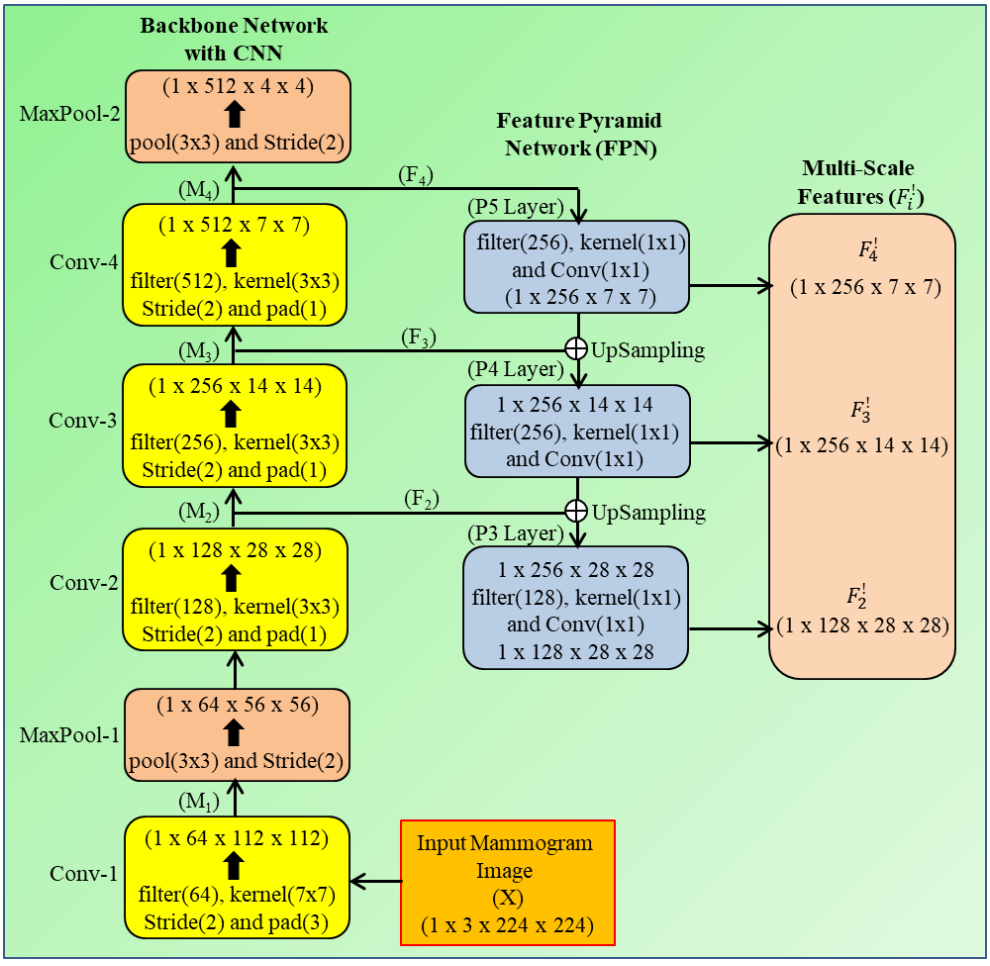

Figure 4.4: CFPN internals (P3/P4/P5 maps, lateral connections) with receptive-field annotations.

(a) Backbone Choice and Feature Extraction

CFPN uses a pre-trained CNN backbone, such as ResNet-50 or VGG-16, because these models have already shown good performance in medical image analysis. The backbone gradually reduces the spatial resolution of the feature maps while increasing their semantic richness. Let the input mammogram image is $'I'$. After it passes through the backbone, we obtain a set of feature maps at different depths as:

$$F_l = \mathcal{T}_l(I), l = 1,2, \dots, L \qquad (4.12)$$

Here, F_l is the feature map at stage l, and $\mathcal{T}_l(\cdot)$ represents the convolutional transformation at that stage. Shallow levels capture edges and simple textures, while deeper levels encode more abstract tumour and tissue patterns.

Table 4.4 focuses on the multi-scale feature maps produced by the CFPN (levels P3, P4 and P5). It reports their spatial size and number of channels for a 224 × 224 input, and it briefly explains the clinical role of each level. The lowest level F_2^i (P3) keeps fine-scale tissue textures and tiny suspicious regions, F_3^i (P4) balances local details with contextual tissue around typical masses, while the highest level, F_4^i (P5) encodes coarse semantic structures related to global breast shape and large tumours.

Table 4.4: Properties of CFPN Feature Maps at Different Scales

FPN level	**Multi-scale feature map**	**Size (C × H × W) for 224 × 224 input**	**Main role in mammogram analysis**
P3	F^i_2	128 × 28 × 28	Fine-scale tissue texture and small suspicious regions; helps for micro-lesions and subtle edge details.
P4	F^i_3	256 × 14 × 14	Medium-scale structures with more context; useful for typical mass lesions and architectural distortion.
P5	F^i_4	256 × 7 × 7	Coarse, high-level semantic features; captures global breast shape and large tumour patterns.

(b) Lateral Connections and Top–Down Pathway

Simply using deep features is not enough, because we also need fine spatial details for accurate localization. To balance semantic richness and spatial precision, CFPN

applies 1×1 lateral convolutions on feature maps from different backbone stages to align their channel dimensions. Then, a top–down pathway is built using nearest-neighbor upsampling and convolution:

$$P_l = \mathrm{Conv}(F_l + \mathrm{Up}(P_{l+1})) \qquad (4.13)$$

Where P_l is the pyramid feature at level l, $\mathrm{Up}(\cdot)$ resizes higher-level features to match the spatial size of lower-level ones, and the convolution has trainable weights. In this way, high-level semantic information from deep layers flows down and reinforces shallow layers, so local fine details are enriched with global context [58].

(c) Scale Weighting and Aggregation

In mammograms, not all scales are equally important. Very fine scales help to pick up micro-calcifications, while coarser scales are better for large masses and architectural distortion [75]. To adapt to this, CFPN uses a scale attention mechanism. For each pyramid level P_l, a normalised weight α_l is learned:

$$F_{\mathrm{multi}} = \sum_l \alpha_l P_l \,, \sum_l \alpha_l = 1 \qquad (4.14)$$

Here, α_l are attention weights learned during training. They allow the network to up-weight the scales that are more discriminative for a given case, either favouring small-scale detail or large-scale structure [82].

(d) Final Multi-scale Representation

The final multi-scale representation, denoted as F_{CFPN}, is obtained from this weighted combination of pyramid features:

$$F_{\mathrm{CFPN}} = F_{\mathrm{multi}}$$

This feature map is then passed to the Cross-Attention Fusion Module. Because it integrates information from different receptive fields in a consistent way, and it helps MsFuNet cope with the diverse morphology of breast tumours. Table 4.5 gives a layer-wise view of the backbone CNN and CFPN configuration. It lists each stage from the raw input, through convolution and pooling blocks (Conv-1 to Conv-4, MaxPool-1/2), up to the construction of P3, P4 and P5 and the final multi-scale features. For every stage it specifies the kernel size, stride, padding, number of output channels and resulting feature-map size. Table 4.5 clarifies the exact tensor

dimensions to reproduce the implementation or to extend the CFPN for other medical imaging tasks.

Table 4.5: Backbone CNN and CFPN configuration for 224 × 224 mammogram input

Stage ID	Block / layer	Main operation (kernel, stride, padding)	Output channels	Output feature size (C × H × W)
S0	Input image	–	3	3 × 224 × 224
S1	Conv-1	Conv, 7×7, stride 2, pad 3	64	64 × 112 × 112
S2	MaxPool-1	Max-pool, 3×3, stride 2	64	64 × 56 × 56
S3	Conv-2 (→ M_2)	Conv, 3×3, stride 2, pad 1	128	128 × 28 × 28
S4	Conv-3 (→ M_3)	Conv, 3×3, stride 2, pad 1	256	256 × 14 × 14
S5	Conv-4 (→ M_4)	Conv, 3×3, stride 2, pad 1	512	512 × 7 × 7
S6	MaxPool-2	Max-pool, 3×3, stride 2	512	512 × 4 × 4
S7	P3 layer	1×1 conv on M_2	128	128 × 28 × 28
S8	P4 layer	1×1 conv on M_3 + upsampled P5	256	256 × 14 × 14
S9	P5 layer	1×1 conv on M_4	256	256 × 7 × 7
S10	Multi-scale F^i_2	From P3	128	128 × 28 × 28
S11	Multi-scale F^i_3	From P4	256	256 × 14 × 14
S12	Multi-scale F^i_4	From P5	256	256 × 7 × 7

4.2.4 Cross-Attention Fusion Module

While CFPN gives us good multi-scale features, we still need a smart way to combine them. If we simply concatenate or sum them, small but important details may be lost. For this reason, MsFuNet adds a Cross-Attention Fusion Module (Figure 4.5), which selectively merges the local fine-scale cues with global contextual information. This helps the model remain sensitive to scattered micro-lesions and, at the same time, respect the broader tissue patterns [24].

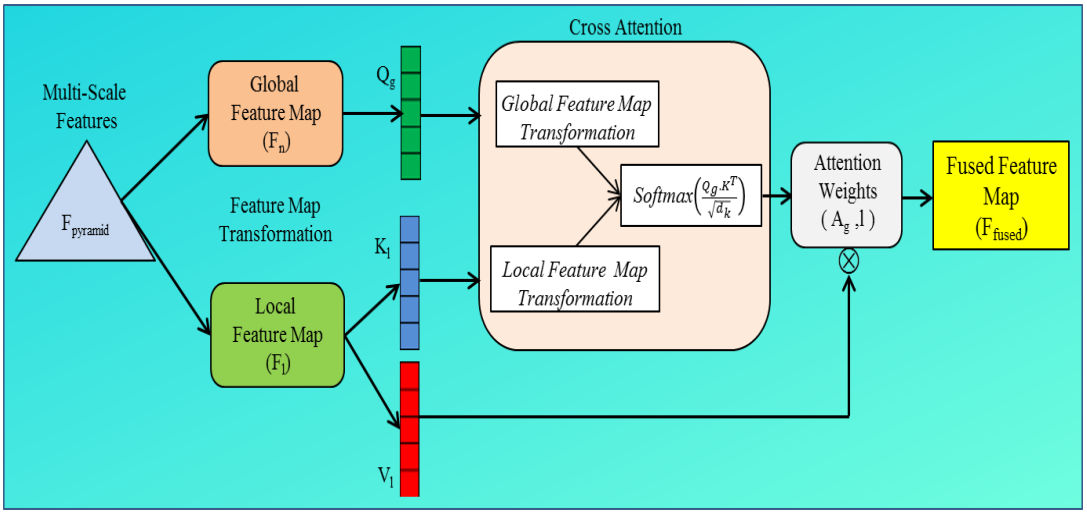

Figure 4.5: Cross-attention fusion (global context attends local features).

(a) Query, Key, and Value Formation: From the CFPN, we separate the feature maps into two streams (see Table 4.6):

- A global stream F_n, taken from deeper pyramid levels (e.g., P5), which carries semantic-rich, low-resolution context about anatomical structure;
- A local streamF_l, taken from shallower levels (e.g., P3), which preserves high-resolution details such as microcalcifications and sharp boundaries.

These are projected into query (Q), key (K), and value (V) tensors using learnable linear layers:

$$Q = W_Q F_n, K = W_K F_l, V = W_V F_l \qquad (4.15)$$

Where W_Q, W_K, W_V are trainable projection matrices. The idea is that global context asks questions (queries) and local detail provides answers (keys and values).

(b) Attention Weighting: Next, cross-attention is computed by matching global queries with local keys:

$$A = \text{softmax}(\frac{QK^\top}{\sqrt{d_k}}) \qquad (4.16)$$

where d_k is the dimensionality of the key vectors. The softmax normalisation ensures that the attention scores are stable and do not grow too large. This step tells the model which local locations are most relevant in the context of the global anatomical pattern.

(c) Fused Feature Computation: The final fused feature map is obtained by applying the attention weights to the value vectors:

$$F_{\text{fused}} = AV \qquad (4.17)$$

Here, local features are enhanced or suppressed depending on how important they are with respect to the global context [83]. For example, scattered micro-calcifications inside dense tissue will receive stronger emphasis if they align with globally suspicious patterns, rather than being treated as random noise.

(d) Intuition for Mammographic Relevance: The motivation for using cross-attention in mammogram analysis is quite intuitive:

- Global context helps avoid false positives. It ensures that small bright spots are interpreted in a realistic anatomical setting, rather than as isolated artifacts.
- Local salience makes sure that small but clinically important details are not ignored, especially in dense or heterogeneous breasts.

By combining these two streams, the module tries to balance specificity (not over-calling noise) and sensitivity (not missing true early lesions).

4.2.5 Domain-Agnostic Regularization Strategy

One of the main problems in mammogram analysis is that different datasets look quite different. Images may vary in resolution, contrast, scanner type, and even patient population. If we train a deep model directly on one dataset, it usually learns many dataset-specific patterns and then performs poorly when it sees images from a new source [22].

To reduce this issue, MsFuNet uses a Domain-Agnostic Regularization Strategy. This strategy combines several ideas: domain-adversarial learning, feature-space augmentation, dropout, weight decay, and domain-specific batch normalization [17]. The overall aim is to make the learned features less sensitive to domain-specific noise, while still keeping them discriminative enough for tumour detection [75].

(a) Domain-Adversarial Learning with Gradient Reversal: The central component is a Domain-Adversarial Neural Network (DANN) training scheme, implemented with a Gradient Reversal Layer (GRL). The shared feature extractor

produces an embedding Fthat is used both by the main heads (segmentation and classification) and by a domain discriminator $D(\cdot)$. The discriminator tries to predict from which dataset a feature comes (i.e., DDSM vs. INbreast) [84]. The domain discriminator is trained with a standard classification loss as:

$$L_{dom} = -\frac{1}{N}\sum_{i=1}^{N}\left[d_i \log \dot{d}_i + (1 - d_i)\log\left(1 - \dot{d}_i\right)\right] \quad (4.18)$$

where the true domain label and the predicted domain probability appear. During back propagation, the GRL multiplies the gradient by a negative factor

$$\frac{\partial L_{dom}}{\partial F} \leftarrow -\vartheta \frac{\partial L_{dom}}{\partial F} \quad (4.19)$$

In practice, this means:

- The discriminator is trained to better separate domains,
- While the feature extractor is forced to fool the discriminator by making features that look similar across domains.

As a result, the shared representation becomes more domain-invariant, which helps in cross-dataset generalization [19].

(b) Feature-Space Augmentation: In addition to the image-level augmentation, MsFuNet also perturbs features in the latent space. A small, random perturbation is added to the feature map F as:

$$F' = F + \epsilon \quad (4.20)$$

where ϵ controls the strength of the perturbation. By slightly "shaking" the feature representation, the model sees a wider variety of feature configurations and is less likely to overfit to narrow patterns from a single dataset.

(c) Dropout and Weight Decay: Inside the network, we also apply standard regularization methods: Dropout randomly turns off a subset of neurons during training as:

$$h' = m \odot h, \quad m \sim Bernoulli(p), \quad (4.21)$$

where h is the neuron activation and m is a binary mask with dropout probability p. This reduces co-adaptation between units and improves robustness. Weight decay

(L2 regularization) adds a penalty proportional to the squared norm of the weights as:

$$L_{wd} = \beta \sum_{l=1}^{L} \|W_l\|_2^2 \qquad (4.22)$$

with β controlling how strongly large weights are penalized. This encourages smoother and simpler models that generalize better.

(d) Domain-Specific Batch Normalization: Because of different datasets can have different intensity distributions, using a single set of batch-normalization statistics may not be ideal. To address this, MsFuNet uses domain-specific batch normalization. For each domain d, separate mean and variance μ_d and σ_d^2 are maintained

$$\dot{x}_d = \frac{x-\mu_d}{\sqrt{\sigma_d^2 + \epsilon}}, \; y = \gamma x_d + \beta, \qquad (4.23)$$

At the same time, shared learnable scale and shift parameters γ and β are used. In this way, each domain retains its own normalization statistics, but the network still learns a common transformation on top. This helps to stabilize training when data come from mixed sources [2 and 8].

(e) Joint Regularization Objective: All these regularization elements contribute to the domain regularization term in the final training objective

$$L_{total} = \vartheta_1 L_{seg} + \vartheta_2 L_{cls} + \vartheta_3 L_{dom} + \vartheta_4 L_{wd} \qquad (4.24).$$

This term is combined with the segmentation and classification losses, so that:

- the model learns good task performance, and
- at the same time reduces sensitivity to domain differences.

This integrated design helps MsFuNet to work more reliably when it is applied to images from new hospitals or screening centres.

4.2.6 Output Heads and Post-processing

MsFuNet uses two output heads on top of the shared backbone and attention fusion: one for pixel-wise segmentation and one for image/region-level classification [85]. This combination provides both precise localization and an overall decision about malignancy.

(a) Segmentation Head: The segmentation branch takes the fused feature map and produces a probability map for each pixel as:

$$\dot{y}_{seg} = \sigma\left(W_{seg} * F_{fused} + b_{seg}\right) \quad (4.25).$$

Depending on the task, a sigmoid activation is used for binary segmentation or a softmax for multi-class segmentation. To get a binary mask, a fixed threshold τ (usually 0.5) is applied to the probability map as:

$$M(x, y) = 1_{\{y_{seg}(x,y) > \tau\}} \quad (4.26).$$

Pixels with probability above τ are marked as tumour, and others as background. To further clean the mask, simple morphological operations are carried out, such as: Erosion and dilation, for removal of tiny isolated components, and contour refinement [15]. Due to this, the small noisy regions are removed and tumour boundaries become smoother and more anatomically plausible.

(b) Classification Head: For lesion-level or image-level classification, the fused feature map is processed by Global Average Pooling followed by fully connected layers

$$\dot{y}_{cls} = softmax\left(W_{cls} \cdot GAP\left(F_{fused}\right) + b_{cls}\right) \quad (4.27)).$$

Global Average Pooling GAP$(\cdot)$ summarizes each channel into a single value, which is then passed through one or more dense layers to produce a vector of class probabilities $\hat{y}$. This vector gives the malignancy likelihood (for example, benign vs malignant) for the whole image or for a selected region. The classification head is trained jointly with the segmentation head, so both tasks share the same feature space.

(c) Confidence Scoring: Along with the predicted class, MsFuNet also computes a confidence score

$$Conf(x) = max_j y_{cls}^{(j)} \quad (4.28)).$$

This score reflects how certain the model is about its decision and can be useful for clinical triage. For example, cases with low confidence can be flagged for closer human review, while high-confidence benign predictions may help reduce unnecessary recalls. In this way, the network output becomes not only a classification label but also an interpretable measure of reliability [41].

4.2.7 Optimization and Loss Functions

To train the MsFuNet in a balanced way, we use a composite loss that combines segmentation, classification, and adversarial (domain) components [11] as shown in Table 4.6. Each part focuses on a different aspect of performance, and together they guide the network towards robust and clinically meaningful behaviours.

(a) Segmentation Loss: For segmentation, MsFuNet uses a hybrid Dice–BCE loss as:

$$\mathcal{L}_{\text{seg}} = \alpha\, \mathcal{L}_{\text{Dice}} + (1-\alpha)\, \mathcal{L}_{\text{BCE}}, \qquad (4.29)$$

where $\alpha \in [0,1]$ controls the relative weight. The Dice loss $\mathcal{L}_{\text{Dice}}$ measures the overlap between the predicted mask and the ground truth. It is robust to class imbalance, which is very important for small tumours as:

$$L_{Dice} = 1 - \frac{2\sum_i p_i g_i}{\sum_i p_i + \sum_i g_i + \epsilon}, \quad (4.30)$$

Similarly, the Binary cross-entropy (BCE) $\mathcal{L}_{\text{BCE}}$ [19] which penalizes pixel-wise classification errors and stabilizes the probability estimates calculated as:

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[g_i\, log(p_i) + (1-g_i)\, log\,(1-p_i)] \quad (4.31)$$

The combination ensures that both region overlap and pixel-level accuracy are optimized.

(b) Classification Loss: The classification head is trained using categorical cross-entropy

$$\mathcal{L}_{\text{cls}} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c), \qquad (4.32)$$

where C is the number of classes, y_c is the true label (one-hot encoded), and $\hat{y}_c$ is the predicted probability for class c. This loss encourages the model to assign high probability to the correct category and low probability to others.

(c) Adversarial (Domain) Loss: The domain discriminator is trained with a standard cross-entropy loss for domain classification

$$L_{dom} = -\frac{1}{N}\sum_{i=1}^{N}\left[d_i \, log\dot{d}_i + (1 - d_i) \, log\left(1 - \dot{d}_i\right)\right] \quad (4.33).$$

This adversarial term, together with the GRL, pushes the feature extractor to create domain-invariant features. In the overall training, the discriminator tries to minimize this loss [22], while the feature extractor tries to maximize it (through gradient reversal), creating an adversarial game that supports domain generalization.

(d) Final Joint Objective: The final training objective is a weighted sum of all components:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{seg}}\mathcal{L}_{\text{seg}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{dom}}\mathcal{L}_{\text{dom}} + \lambda_{\text{wd}}\mathcal{L}_{\text{wd}}, \qquad (4.34)$$

where $\lambda_{\text{seg}}, \lambda_{\text{cls}}, \lambda_{\text{dom}}, \lambda_{\text{wd}}$ are hyperparameters that control the strength of each term, and $\mathcal{L}_{\text{wd}}$ represents the weight decay regularization. These weights are tuned so that the segmentation quality remains high, classification is reliable, and domain invariance and generalization are encouraged, without any single component dominating the others.

Table 4.6: Loss Functions, Symbols, and Roles in MsFuNet

Loss	Equation	Symbols	Role
Dice Loss (L_{Dice})	$1 - \frac{2\sum p_i g_i}{\sum p_i + \sum g_i}$	p_i: predicted prob, g_i: ground truth	Measures overlap, robust to imbalance

BCE Loss (L_{BCE})	$-\frac{1}{N}\sum g\ log\ p + (1-g)\ log\ (1-p)$	N: pixels, p: prob	Pixel-wise error minimization
Classification CE (L_{cls})	$-\sum y_j\ log\ \dot{y}_j$	y_j: true label, $\dot{y}_j$: predicted	Ensures correct class discrimination
Domain Loss L_{dom}	$-\sum d\ log\ \dot{d}$	d: domain label	Enforces domain invariance
Weight Decay (L_{wd})	$\beta \sum \|W\|^2$	W: weights	Prevents overfitting, smooths training

Finally, this optimization setup lets MsFuNet learn rich, shared features that support both segmentation and classification, while remaining robust across datasets and less sensitive to domain-specific artifacts.

4.3 Experimental Setup

4.3.1 Datasets and Splits

The experimental study for MsFuNet was carried out on three well-known mammography datasets: DDSM, INbreast, and MIAS [76]. These datasets were chosen because together they cover a wide range of imaging conditions (See figure 4.6), from digitized film mammograms to modern full-field digital mammography (FFDM). In addition, they also differ in resolution, lesion types, and annotation style [86]. This combination is useful to test how well MsFuNet can handle domain shift in practice.

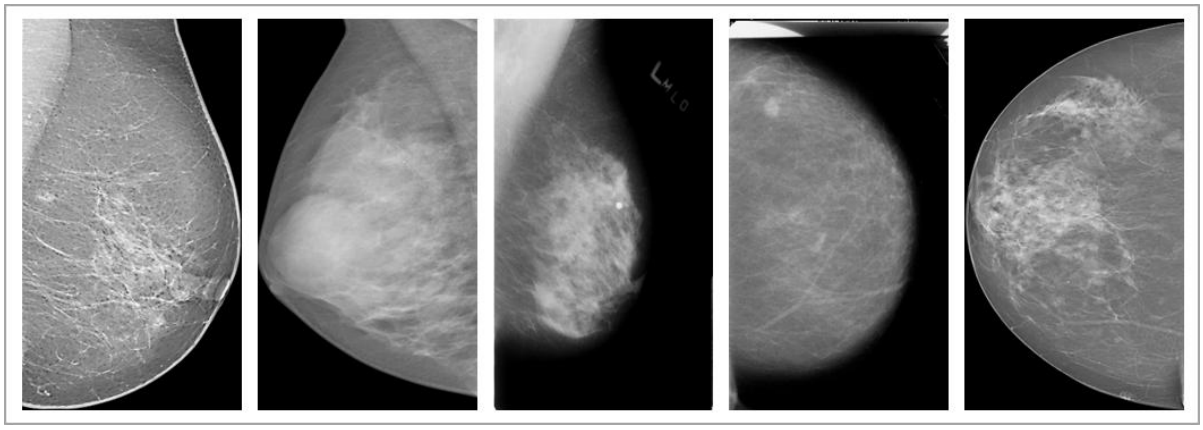

Figure 4.6 Sample Mammogram images from DDSM, INBreast and MIAS datasets

(a) DDSM Dataset

The Digital Database for Screening Mammography (DDSM) is one of the most widely used film-based mammogram collections. It contains around 2,620 cases and more than 10,000 images, which are scanned from film using HOWTEK and LUMYSIS scanners with spatial resolutions between 50 μm and 200 μm. For each case, two standard views per breast are available, together with metadata such as:

- BI-RADS breast density,
- Subtlety ratings, and
- Textual descriptions of each abnormality.

Pixel-level masks are provided for both benign and malignant lesions, so the dataset is suitable for both segmentation and classification tasks [87]. Because the images come from different scanners and scanning settings, DDSM shows clear variability in noise level and contrast. This makes it a good test bed for domain generalization.

(b) INbreast Dataset

The INbreast dataset is a high-quality FFDM collection containing 115 cases and 410 images, with resolutions up to 3328 × 4084 pixels. It includes several lesion categories, such as:

- Masses,
- Micro-calcifications,
- Asymmetries, and architectural distortions.

All lesions are annotated with precise contours drawn by expert radiologists. The dataset also contains normal and post-mastectomy cases, which increases its diversity. Because of its high resolution and accurate labels, INbreast is especially helpful to check how well MsFuNet can segment fine structures (for example, small clusters of micro-calcifications). However, this high resolution also increases memory and computation demands during training and testing.

(c) MIAS Dataset

The MIAS (Mammographic Image Analysis Society) dataset is a classical benchmark with 161 cases and 322 digitized images, each stored at a resolution of 1024 × 1024 pixels. It includes a variety of findings such as:

- Spiculated masses,
- Calcifications,
- Asymmetries, and architectural distortions.

Unlike INbreast, MIAS do not provide full pixel-level masks. Instead, lesions are usually annotated by a centre coordinate and an approximate radius. This makes the segmentation task less precise, but the dataset is still important historically and is widely used for algorithm comparison in mammography.

(d) Cross-Validation Strategy

To obtain reliable and unbiased results, a 3-fold cross-validation (CV) strategy was used for each dataset. The splitting was done at case level, not at image level. In every fold, the cases were divided as:

- 70% for training,
- 10% for validation, and
- 20% for testing [88].

By splitting at case level, both breasts of the same patient always appear in the same subset. This avoids data leakage between train and test sets (Table 4.7). For each metric, the final performance is reported as the average across the three folds, as indicated

$$Metric_{Avg} = \frac{1}{k} \sum_{k=1}^{k} Metric^{(k)} \qquad (4.35),$$

where $K = 3$ the number of folds and the result of fold is $k$is averaged over all folds.

Table 4.7: Dataset Statistics and Annotation Characteristics

Dataset	Cases	Images	Resolution / Scale	Lesion Types	Annotation Type

DDSM	2,620	10,480	50–200 μm (digitized film)	Masses, calcifications, benign / malignant findings	Pixel-level contours
INbreast	115	410	Up to 3328 × 4084 px	Masses, calcifications, distortions, asymmetries	Precise expert contours
MIAS	161	322	1024 × 1024 px	Masses, calcifications, asymmetries	Approx. centre & radius

4.3.2 Implementation Details

(a) Framework and Hardware: All experiments were implemented in PyTorch v1.10, with NumPy, SciPy, and OpenCV used for data loading, preprocessing, and augmentation. Training was mainly done on a single NVIDIA Tesla V100 GPU (32 GB), which provides enough memory for 512 × 512 inputs and mixed-precision training. For completeness, inference was also checked on an Intel Xeon CPU (64 cores) to confirm that the model can run, although much slower, in a CPU-only environment [27]. Mixed-precision training (FP16/FP32) using NVIDIA Apex was enabled to reduce memory usage and speed up training without noticeable loss in accuracy.

(b) Seed Control and Reproducibility: To make the experiments reproducible, we fixed the random seeds for all major libraries (Python, NumPy, and PyTorch), as summarized below:

$$seed_{torch} = seed_{numpy} = seed_{random} = 42. \qquad (4.36).$$

By doing this, the same sequence of random numbers is used across runs, which keeps: Mini-batch shuffling, Data augmentation operations, and weight initialization as close as possible between repeated experiments. This is important for a fair comparison of different model variants and for checking the stability of MsFuNet.

(c) Training Protocol: For training of our model, the following training setup was used.

- **Batch size and number of epochs**

- Batch size was set to 16 for input size 512 × 512.
- For high-resolution INbreast experiments, the batch size was reduced to 8 due to memory limits.
- Maximum number of epochs was 100, with early stopping if the validation loss did not improve for 15 consecutive epochs.

- **Optimizer and learning rate schedule:** Training used the Adam optimizer with default momentum parameters$(\beta_1, \beta_2) = (0.9,0.999)$. The update rule follows the standard Adam formulation

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)gt,\ v_t = \beta_2 v_2 + (1-\beta_2)g_t^2,\ \theta_{t+1} = \theta_t - n \cdot \frac{m_t}{\sqrt{v_t + \epsilon}} \text{ (4.37)),}$$

- where the learning rate η is applied to the bias-corrected first and second moment estimates of the gradient at step t.

The learning rate followed a step-decay schedule:

- Initial learning rate $\eta_0$was set to 0.001 by default.
- Every 30 epochs, the learning rate was multiplied by a factor $\gamma = 0.1$, as

$$\eta_t = \eta_0 \cdot \gamma^{\left[\frac{t}{T}\right]} \quad . \text{(4.38).}$$

This scheme allows faster learning at the beginning and more careful refinement later in training.

- **Regularization**
  - Dropout with rate 0.5 was applied in the decoder and classifier layers to reduce overfitting.
  - Weight decay (L2 regularization) was set to 1×10^{-4}, which penalizes large weights and encourages smoother solutions.
  - Mixed-precision training further helps to stabilize gradients and speeds up training.

(d) Training Hyperparameters: Table 4.8 summarizes the main hyperparameters and indicates the default values (from initial pilot runs) and the final tuned values used in the reported experiments.

Table 4.8: Training Hyperparameters and Schedules

Hyperparameter	Default Value	Tuned Value	Notes / Remarks
Initial LR	0.001	0.0005	Smaller LR for stable training on INbreast
Optimizer	Adam	Adam	(\beta_1 = 0.9), (\beta_2 = 0.999)
Batch Size	32	16 (8 for INbreast)	Limited by GPU memory for high-res inputs
Epochs	50	100	With early stopping (patience = 15)
Dropout Rate	0.3	0.5	Stronger regularization against overfitting
Weight Decay	5×10^{-5}	1×10^{-4}	Slightly stronger L2 penalty
LR Decay Step	20	30	Slower decay for smoother convergence

In total, this experimental setup was chosen to balance training stability, computational cost, and fair comparison across datasets. It also matches realistic constraints for deploying such a model on typical research or hospital GPU hardware.

4.3.3 Augmentation Protocol

Because mammogram datasets are relatively small and also quite heterogeneous, data augmentation is very important to improve the robustness of MsFuNet [5]. In this work, we use a dedicated augmentation pipeline that is slightly adapted to the imaging properties of DDSM, INbreast, and MIAS. Augmentations are applied randomly during training, so that each mini-batch contains a mixture of original and transformed images (Table 4.9). This helps the model to see a wider variety of appearances without changing the underlying anatomy.

The main transformations are:

- **Rotation:** Random rotations in the range $[-15°, +15°]$are applied to simulate variations in breast positioning during acquisition. The rotation operator is represented in

$$I' = R_\theta (I),\ \theta \sim U(15,15) \quad (4.39).$$

- **Horizontal Flip:** A left–right flip is applied with probability $p = 0.5$. This reflects the approximate bilateral symmetry of the breasts and helps the network not to overfit to a preferred orientation.

- **Intensity Adjustments:** Brightness and contrast are randomly scaled to mimic differences in scanner calibration and patient-specific attenuation. The transformation is expressed in

$$I'' = \gamma I' + \beta, \quad (4.40),$$

where the scaling and shifting factors control contrast and brightness.

- **Gaussian Noise:** Low-level Gaussian noise with a small variance is added to simulate acquisition artifacts, especially relevant for digitized film images like DDSM. This is written in as:

$$I_{arg} = I'' + N(0, \sigma^2) \ (4.41).$$

The overall augmentation strategy has dataset-specific motivation:

- For DDSM, augmentation helps to compensate for inconsistent film digitization and scanner noise.

- For INbreast, the high native resolution allows slightly stronger perturbations (for example, more noticeable noise injection) without destroying fine details.

- For MIAS, which has lower resolution, rotations and contrast changes are more emphasized to make the model less sensitive to limited spatial detail.

Table 4.9: Augmentation Transformations and Probabilities

Transformation	Parameter Range	Probability	Rationale for Use
Rotation	([-15°, +15°])	0.8	Simulates variation in patient positioning

Horizontal Flip	Left–right mirroring	0.5	Exploits bilateral breast symmetry
Intensity Scaling	Random brightness/contrast	0.6	Mimics scanner and contrast differences
Gaussian Noise	Small variance (σ^2)	0.4	Increases robustness to acquisition noise

4.3.4 Evaluation Metrics

To evaluate MsFuNet in a comprehensive way, we use both segmentation metrics and classification metrics [26] as defined in Table 4.10. This allows me to measure not only how well the model localizes lesions, but also how accurately it decides between benign and malignant cases.

(a) Segmentation Metrics

- **Pixel Accuracy (PA):** PA measures the proportion of correctly classified pixels (both tumour and background) over the entire image. It is defined in terms of true positives, true negatives, false positives, and false negatives.

$$PA = \frac{TP+TN}{TP+TN+FP+FN} \text{ (4.42),}$$

- **Intersection-over-Union (IoU):** IoU, also known as the Jaccard Index, is given in

$$IoU = \frac{|P \cap G|}{|P \cup G|} = \frac{TP}{TP+FP+FN} \text{ (4.43).}$$

  It measures the overlap between the predicted mask $P$and the ground truth $G$by dividing the intersection area by the union area.

- **Dice Coefficient (DSC):** The Dice coefficient is another overlap measure that is particularly useful when the lesion region is small. It gives more weight to agreement in the tumour area and is less affected by background dominance.

$$DSC = \frac{2TP}{2TP+FP+FN} \text{ (4.44)}$$

(b) Classification Metrics

Accuracy (ACC): Overall classification accuracy is the fraction of correctly classified images or regions among all samples.

$$ACC = \frac{TP+TN}{TP+TN+FP+FN}. \quad (4.45)$$

- **Precision (PRC):** Precision reflects the fraction of predicted positive cases that are actually true positives. High precision means fewer false alarms.

$$PRC = \frac{TP}{TP+FP} \quad (4.46)$$

- **Recall (Sensitivity, SE):** Recall measures how many of the true positive cases are correctly detected. High recall is important in screening to avoid missed cancers.

$$SE = \frac{TP}{TP+FN} \quad (4.47)$$

- **F1-Score:** The F1-score is the harmonic mean of precision and recall. It is helpful when classes are imbalanced, because it penalizes models that sacrifice one metric for the other.

$$F1 = \frac{2 \cdot PR \cdot SE}{PR+SE} \quad (4.48)$$

- **Area under the ROC Curve (AUC):** AUC summarizes the performance of the classifier across all possible decision thresholds. It gives a threshold-independent view of discrimination ability.

$$AUC = \int_0^1 TPR(FPR)dFPR. \quad (4.49)$$

Table 4.10: Segmentation and Classification Metric Definitions and Roles

Metric	Formula	Role in Evaluation
PA	$\frac{TP+TN}{TP+TN+FP+FN}$	Pixel-wise correctness
IoU	$\frac{TP}{TP+FP+FN}$	Overlap accuracy
Dice	$\frac{2TP}{2TP+FP+FN}$	Robust to small lesions
Precision	$\frac{TP}{TP+FP}$	Positive prediction quality

Recall	$\frac{TP}{TP + FN}$	Sensitivity to true lesions
F1	$\frac{2 \cdot PR \cdot SE}{PR + SE}$	Balanced performance
AUC	$\int_0^1 TPR(FPR)dFPR$	Threshold-free classification quality

4.3.5 Baselines and Comparators

To properly judge how strong MsFuNet is, it is compared against a set of well-known state-of-the-art baseline models [23]. These baselines are divided into two families (Table 5.11) are: segmentation models and classification models. All models were re-implemented or carefully adapted using a consistent training setup to keep the comparison fair.

(a) Segmentation Baselines

- **U-Net** – A classical encoder–decoder architecture with skip connections, widely used in medical image segmentation.
- **U-Net++** – An extension of U-Net with nested skip connections, designed to capture more fine-grained details.
- **TransUNet** – Combines a CNN encoder with a Transformer-based decoder to use both local and global features.
- **Attention U-Net** – Adds attention gates to the U-Net framework, so that the network focuses more on relevant regions.
- **ResUNet** – Incorporates residual blocks into the U-Net style architecture to stabilize gradient flow.
- **D-UNet** – Uses dense connections to improve feature reuse and encourage richer representations.
- **MammoUNet** – A U-Net variant specifically adapted for mammography images.

(b) Classification Baselines

- **EfficientNetV2** – A family of convolutional models found by neural architecture search, with optimized scaling.

- **DenseNet201** – Uses dense connectivity between layers to maximize feature reuse.
- **Vision Transformer (ViT)** – Applies self-attention to image patches, treating images as sequences of tokens.
- **ResNet-50** – A 50-layer residual network, widely used as a backbone in many vision tasks.
- **Swin Transformer** – A hierarchical vision transformer with shifted windows, capturing both local and global structure see Table 4.11.

Table 4.11: Baseline Model Configurations

Task	Input Size	Model	Params (M)	Pretrained?
Segmentation	512 × 512	U-Net	7.8	No
		U-Net++	9.2	No
		TransUNet	105	Yes (ImageNet)
		Attention U-Net	8.5	No
Classification	224 × 224	DenseNet201	20	Yes (ImageNet)
		EfficientNetV2	13.6	Yes (ImageNet)
		ResNet-50	25.6	Yes (ImageNet)
		ViT-B/16	86	Yes (ImageNet-21k)
		Swin Transformer	88	Yes (ImageNet-21k)

4.3.6 Statistical Testing and Reporting

To make sure that the improvements of MsFuNet over the baselines are not just due to random variation, we followed a simple but careful statistical testing protocol.

- **Cross-Validation:** Each dataset was evaluated using k-fold cross-validation with $k = 3$. For every metric, to report the mean and standard deviation across the three folds. This reduces the risk that the results are biased by a particular train–test split [14].
- **Paired Significance Tests:** For metrics that appeared approximately normally distributed, we used the paired t-test to compare MsFuNet with each baseline. The test checks whether the average difference in performance across folds is significantly different from zero. When the normality

assumption was not clearly satisfied, applied the Wilcoxon signed-rank test, which is a non-parametric alternative.

- **Effect Sizes:** In addition to p-values, computed Cohen's d

$$d = \frac{\mu_1 - \mu_2}{s_p}, \quad s_p = \sqrt{\frac{(s_1^2 + s_2^2)}{2}} \quad (4.50))$$

- To quantify the magnitude of the difference between MsFuNet and each baseline. Cohen's d uses the means and standard deviations of the two methods to provide a more interpretable effect size.

- **Confidence Intervals (CI):** For the main metrics, 95% confidence intervals were obtained by bootstrapping with 1,000 resamples. This gives an idea of the uncertainty around the estimated performance.

- **Multiple-Comparison Control:** Because many metrics and many baselines are tested, the risk of Type I error (false positives) increases. To reduce this, the Holm–Bonferroni correction is used across the set of hypothesis tests [89].

Taken together, these steps ensure that the observed gains of MsFuNet are statistically reliable and also clinically meaningful, rather than artifacts of a particular data split or random fluctuation.

4.4 Results

4.4.1 Training Dynamics

In the first step, we analysed the learning behaviours of the MsFuNet during the training and compare, it with the baseline models. This analysis is essential, for the training curves are provide insight not only into the final performance but also into the speed and stability with each model converges. Figure 4.7 illustrates the evolution of three segmentation metrics are—Pixel Accuracy (PA), Intersection over Union (IoU), and Dice Coefficient (DSC)— across 100 training epochs for the DDSM, INbreast, and MIAS datasets. It provides a temporal perspective on the models' learning behaviours and the offer insights into their convergence the speed and stability.

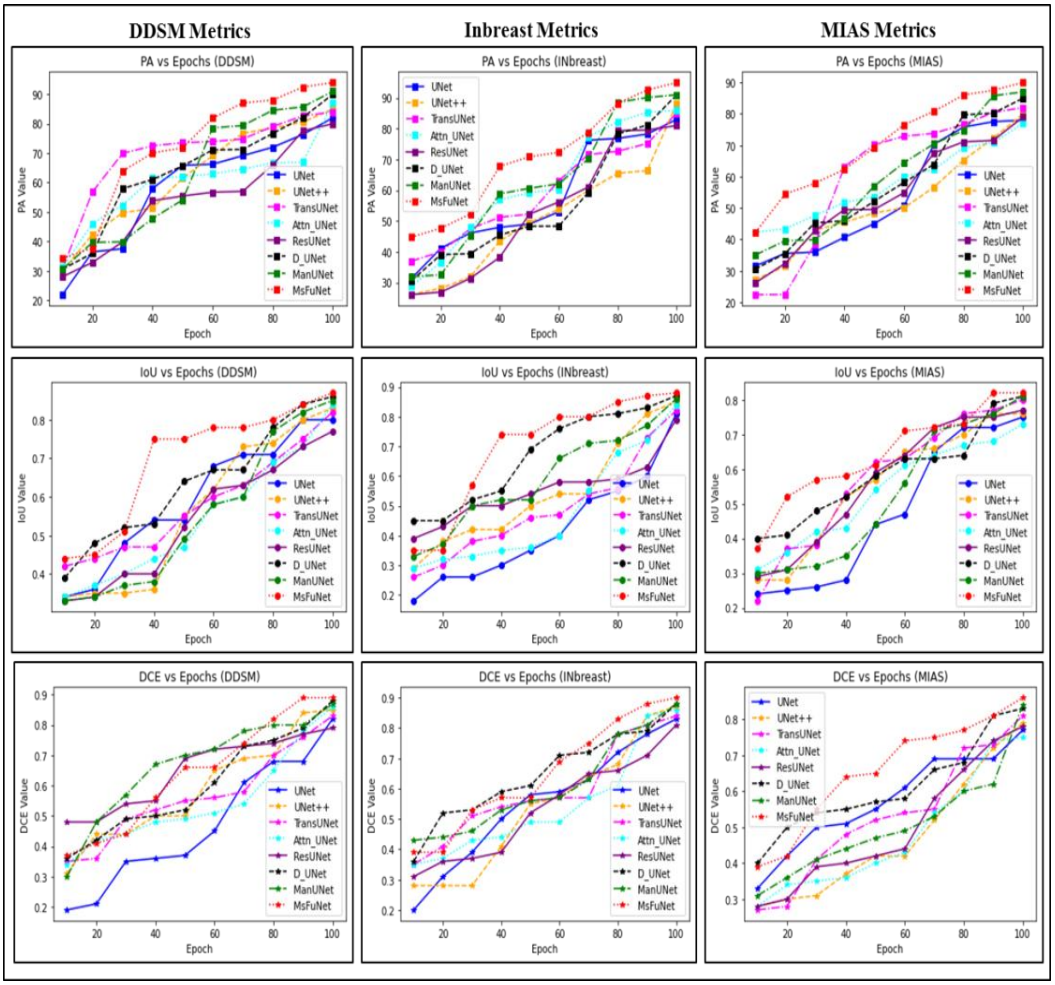

Figure 4.7: Learning curves (PA/IoU/Dice) on DDSM, INbreast, and MIAS datasets across 100 epochs.

From these trajectories, the following points were observe red.

1) Fast Convergence of MsFuNet: Across all the three datasets, MsFuNet attains the high PA, IoU, and Dice values within the first 30–40 epochs, whereas the baseline models are such as U-Net and ResUNet remain in a steep learning phase during this period. This accelerated convergence appears to be primarily attributable to the Cross-Attention Fusion Module. This enables the network to integrate the global anatomical context with the local tumour regions features early in the training process [62].

$$M_{(t)}^{d} = f_d(\theta_t), \quad d\epsilon\{PA, IoU, DSC\} \text{ (4.51)},$$

Where the performance gain per epoch is denoted by $\Delta M(t)$. MsFuNet exhibits a steeper $\Delta M(t)$ slope during approximately the first 20 epochs, indicating more

efficient gradients utilization and faster adaptation to the underlying training data distribution [29].

2) Smooth and Stable Improvements: In contrast to the ResUNet, which frequently exhibits the fluctuations and oscillations in IoU, MsFuNet demonstrates smooth and largely monotonic performance improvements. Even on smaller datasets like MIAS, where most baselines even start to overfit, MsFuNet continues to improve gradually and ends with a higher Dice score. This suggests that the combination of CFPN, cross-attention, and domain-agnostic regularization helps to stabilize the learning process.

3) Dataset-Specific Behaviours: On DDSM (the largest dataset), all models improve steadily, but MsFuNet keeps a consistent 2–3% advantage in PA over the whole training period. With INbreast, the effect of high-resolution inputs is visible. The MIAS, which has fewer samples and lower resolution, MsFuNet still shows strong learning behaviours under data scarcity. While baselines become flat early, MsFuNet continues to improve even after epoch 50 [30]. These training curves show that MsFuNet is not only able to reach better final performance, but it reaches it faster and more consistently.

4.4.2 Segmentation – In-Domain Tests

Next, in-domain segmentation performance is evaluated, in which training and testing were conducted on the same dataset using 3-fold cross-validation. This setup assesses how the effectively each model is adapting to the specific characteristics of a given dataset. It provides a reliable estimate the upper performance bound under the controlled experimental conditions. For each model and dataset, PA, IoU, and Dice, are reported as mean ± standard deviation across folds as discussed in Table 4.12, 4.13 and 4.14.

(a) Results on DDSM Dataset

Table 4.12: Segmentation results on DDSM dataset (mean ± std).

Model	PA (%)	IoU	Dice
U-Net	82.3 ± 1.5	0.80 ± 0.07	0.82 ± 0.06
UNet++	85.6 ± 1.4	0.83 ± 0.04	0.85 ± 0.02
TransUNet	84.1 ± 1.3	0.82 ± 0.02	0.83 ± 0.04
Attention U-Net	87.2 ± 1.2	0.84 ± 0.02	0.86 ± 0.02

ResUNet	80.5 ± 1.7	0.77 ± 0.05	0.79 ± 0.03
D-UNet	90.5 ± 1.1	0.86 ± 0.03	0.88 ± 0.02
MammoUNet	91.2 ± 1.0	0.85 ± 0.01	0.87 ± 0.02
MsFuNet	**94.5 ± 0.9**	**0.87 ± 0.02**	**0.89 ± 0.03**

Interpretation:

- From Table 4.12, MsFuNet reaches 94.5% PA, which is about 3.3% higher than the second-best model (MammoUNet).
- A Dice of 0.89 shows that MsFuNet segments tumour regions with better overlap precision on film-based mammograms with varying breast densities.
- ResUNet is the weakest performer in this setting, indicating that its architectural design is less effective at capturing fine structural details and complex patterns within noisy and heterogeneous DDSM images.

(b) Results on INbreast Dataset

Table 4.13: Segmentation results on INbreast dataset (mean ± std).

Model	PA (%)	IoU	Dice
U-Net	83.5 ± 1.5	0.81 ± 0.05	0.83 ± 0.04
UNet++	88.5 ± 1.2	0.85 ± 0.02	0.87 ± 0.03
TransUNet	85.0 ± 1.4	0.82 ± 0.02	0.84 ± 0.02
Attention U-Net	86.5 ± 1.3	0.84 ± 0.02	0.86 ± 0.02
ResUNet	81.7 ± 1.7	0.79 ± 0.06	0.81 ± 0.05
D-UNet	91.0 ± 1.1	0.87 ± 0.01	0.88 ± 0.01
MammoUNet	91.5 ± 1.2	0.86 ± 0.03	0.88 ± 0.03
MsFuNet	**95.0 ± 0.6**	**0.88 ± 0.02**	**0.90 ± 0.02**

Interpretation:

- On this high-quality FFDM dataset, MsFuNet obtains 95% PA, making it the most accurate model (Table 4.13).
- The Dice score of 0.90 indicates very good boundary alignment, including small micro-calcifications and fine-edged lesions.

- Although D-UNet and MammoUNet are competitive, they still do not reach the same boundary precision and robustness as MsFuNet, especially when lesions are subtle and complex.

(c) Results on MIAS Dataset

Table 4.14: Segmentation results on MIAS dataset (mean ± std).

Model	PA (%)	IoU	Dice
U-Net	78.0 ± 1.6	0.75 ± 0.05	0.77 ± 0.04
UNet++	80.0 ± 1.5	0.76 ± 0.03	0.79 ± 0.03
TransUNet	82.0 ± 1.4	0.80 ± 0.02	0.81 ± 0.02
Attention U-Net	77.6 ± 1.2	0.73 ± 0.06	0.75 ± 0.04
ResUNet	79.5 ± 1.7	0.77 ± 0.06	0.78 ± 0.05
D-UNet	85.2 ± 1.1	0.81 ± 0.01	0.83 ± 0.03
MammoUNet	87.7 ± 1.0	0.81 ± 0.03	0.84 ± 0.02
MsFuNet	**90.2 ± 0.9**	**0.82 ± 0.02**	**0.86 ± 0.01**

Interpretation:

- Even though MIAS is small and has lower resolution, Table 4.14 shown that MsFuNet still achieves a Dice of 0.86, about 2% higher than MammoUNet.
- Baseline models such as U-Net and Attention U-Net perform much worse, indicating that they struggle to generalize on small, noisy datasets.
- MsFuNet's performance on MIAS suggests that the Domain-Agnostic Regularization Strategy effectively controls overfitting and maintains good segmentation even when labelled data are limited.

(d) Average Performance across Datasets

To get an overall view, we calculated the average PA, IoU, and Dice across all three datasets (Figure 4.8) using

$$\bar{M} = \frac{1}{3}\sum_{d\epsilon\{DDSM,INbreast,MIAS\}} M_d, \quad (4.52),$$

where $\bar{M}$ denotes the average metric and M_d is the metric on dataset d. Here MsFuNet achieves PA = 93.2%, IoU = 0.86, and Dice = 0.88, which are higher than all other models by a clear margin.

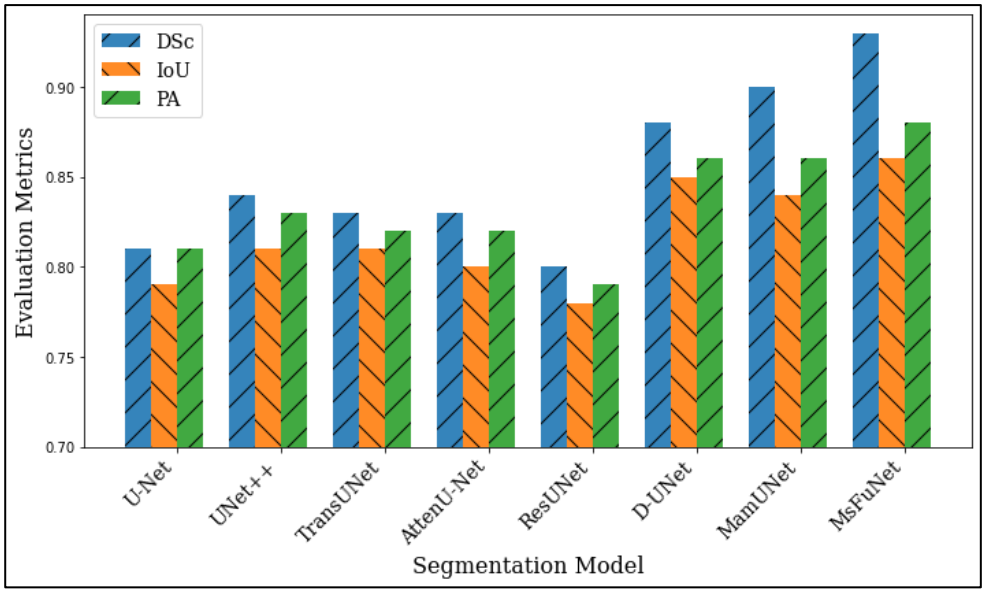

Figure 4.8: Average segmentation performance across DDSM, INbreast, and MIAS with 95% confidence intervals.

This combined view in Figure 4.8 shows that MsFuNet can maintain strong performance on large, high-resolution, and small/limited datasets, which is valuable when moving between different clinical environments.

4.4.3 Classification – In-Domain Tests

Besides segmentation, we also evaluated the MsFuNet for classification, where the goal is to distinguish benign, malignant, and (for some datasets) normal cases. Here the focus moves from pixel-wise details to global image- or ROI-level patterns [4]. By evaluating classification together with segmentation, we can see how well MsFuNet performs as a unified multitask model compared to strong single-task classifiers such as EfficientNetV2, DenseNet201, Vision Transformer (ViT), ResNet-50, and Swin Transformer. For each dataset, classification was also done in-domain with 3-fold cross-validation as presented in Table 4.15, 4.16 and 4.17. The metrics used are Precision, Recall, F1-Score, Accuracy, and AUC-ROC.

(a) DDSM Classification Results: For DDSM, the classification task is binary (benign vs malignant).

Table 4.15: DDSM classification results (mean ± std).

Model	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)	AUC

EfficientNetV2	86.3 ± 1.4	81.4 ± 1.9	83.8 ± 1.6	82.4 ± 1.7	0.87 ± 0.02
DenseNet201	85.9 ± 1.5	83.4 ± 1.7	84.6 ± 1.5	83.3 ± 1.6	0.88 ± 0.02
Vision Transformer	90.1 ± 1.1	89.5 ± 1.2	89.8 ± 1.2	89.5 ± 1.1	0.92 ± 0.01
ResNet-50	87.3 ± 1.3	85.2 ± 1.6	86.3 ± 1.4	85.1 ± 1.5	0.89 ± 0.02
Swin Transformer	91.1 ± 1.0	89.5 ± 1.2	90.3 ± 1.1	89.4 ± 1.3	0.93 ± 0.01
MsFuNet	**94.0 ± 0.8**	**93.5 ± 0.9**	**93.7 ± 0.8**	**93.5 ± 0.8**	**0.96 ± 0.01**

Discussion:

- In Table 4.15, MsFuNet clearly outperforms all baselines, with an F1-Score of 93.7%, which is about 3.4% higher than the next best model (Swin Transformer, 90.3%).
- Analysis indicates fewer false negatives, which is very important clinically because missing a malignant case is more dangerous than an extra false alarm.
- An AUC of 0.96 shows that MsFuNet remains robust when the decision threshold changes, which is useful in different screening settings.

(b) INbreast Classification Results: For INbreast, the classification problem is multiclass (normal, benign, malignant). Due to class imbalance, we report both micro-averaged and macro-averaged metrics in Table 4.16.

Table 4.16: INbreast classification results (multiclass; mean ± std).

Model	**Micro-Prec**	**Micro-Rec**	**Micro-F1**	**Macro-Prec**	**Macro-Rec**	**Macro-F1**	**Acc (%)**	**AUC**
EfficientNetV2	91.4 ± 1.2	89.0 ± 1.3	90.2 ± 1.2	89.2 ± 1.5	87.6 ± 1.6	88.4 ± 1.4	89.3 ± 1.3	0.92 ± 0.01
DenseNet201	90.8 ± 1.3	89.7 ± 1.4	90.2 ± 1.3	88.4 ± 1.6	87.2 ± 1.6	87.8 ± 1.5	88.8 ± 1.5	0.91 ± 0.01
Vision Transformer	92.6 ± 1.1	90.6 ± 1.2	91.6 ± 1.2	90.8 ± 1.3	89.7 ± 1.4	90.1 ± 1.3	90.8 ± 1.2	0.93 ± 0.01
ResNet-50	89.3 ± 1.4	87.5 ± 1.5	88.4 ± 1.4	87.0 ± 1.6	85.9 ± 1.7	86.3 ± 1.6	87.4 ± 1.5	0.90 ± 0.02
Swin Transformer	92.2 ± 1.2	91.9 ± 1.2	92.1 ± 1.2	91.1 ± 1.3	90.3 ± 1.3	90.7 ± 1.3	91.3 ± 1.2	0.94 ± 0.01

MsFuNet	**95.0 ± 0.8**	**94.5 ± 0.9**	**94.7 ± 0.9**	**94.2 ± 0.9**	**93.8 ± 1.0**	**94.0 ± 0.9**	**94.5 ± 0.8**	**0.97 ± 0.01**

Discussion:

- MsFuNet demonstrates the strong multiclass performance, achieving a Macro-F1 score of 94.0%, indicating that it handles all classes are—including the minority categories—in a balanced and consistent manner.
- This improved a class separation can be attributed to the cross-attention mechanism, which are enables the model to focus on subtle discriminative patterns that differentiate malignant lesions from the benign irregularities [90].

(c) MIAS Classification Results: For the MIAS dataset, which is smaller in size and contains lower-resolution images, this experiment evaluates whether the MsFuNet can maintain the strong performance under such constrained conditions is presented in Table 4.17.

Table 4.17 MIAS classification results (mean ± standard deviation).

Model	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)	AUC
EfficientNetV2	86.6 ± 1.5	82.3 ± 1.7	84.4 ± 1.6	83.1 ± 1.7	0.87 ± 0.02
DenseNet201	88.1 ± 1.3	85.2 ± 1.6	86.6 ± 1.4	85.4 ± 1.5	0.89 ± 0.01
Vision Transformer	90.8 ± 1.2	88.1 ± 1.3	89.4 ± 1.2	88.4 ± 1.3	0.91 ± 0.01
ResNet-50	86.5 ± 1.4	85.9 ± 1.6	86.2 ± 1.5	85.8 ± 1.6	0.88 ± 0.02
Swin Transformer	91.4 ± 1.0	88.9 ± 1.2	90.1 ± 1.1	89.2 ± 1.1	0.92 ± 0.01
MsFuNet	**92.5 ± 0.9**	**93.4 ± 0.8**	**92.9 ± 0.9**	**92.1 ± 0.9**	**0.95 ± 0.01**

Discussion:

- Despite the limited dataset size and reduced image resolution, MsFuNet achieves the highest accuracy (92.1%) and AUC (0.95), outperforming with the Swin Transformer by approximately 3%.
- The confusion matrices again reveal that the MsFuNet attains the lowest false-negative rate, an essential to the property for minimizing the risk of undetected malignant lesions.

Observations at Glance: Across the binary (DDSM, MIAS) and multiclass (INbreast) classification tasks, MsFuNet consistently surpasses the strong baseline models [91]. The bar plot in Figure 4.9 shows that MsFuNet spans the largest area across the all evaluated metrics and datasets. This is indicating that not only superior average performance but also a well-balanced performance profile. From a clinical perspective, achieving the high recall while maintaining a low false-negative rates are crucial. The results suggest that the MsFuNet can contribute to safer and more reliable diagnostic decision-making by reducing the likelihood of missed cancers [92].

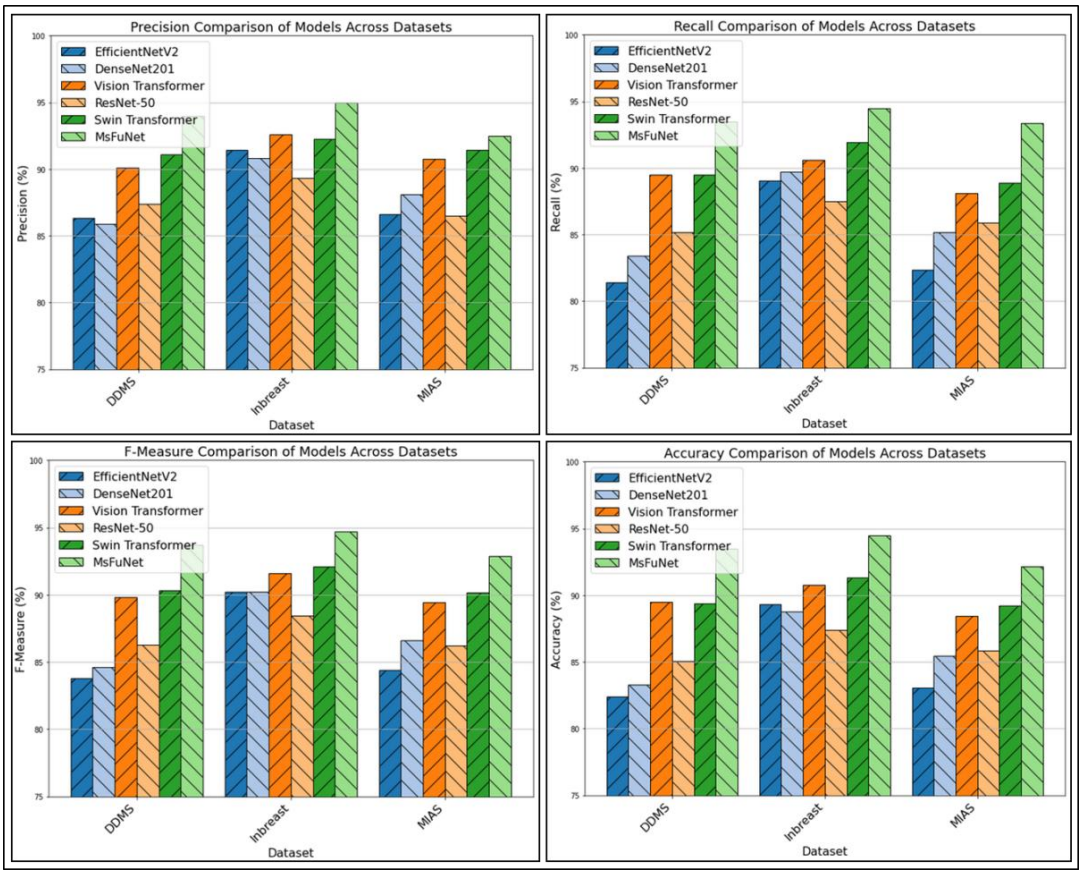

Figure 4.9: Classification Performance Analysis of Models across Multiple Datasets

4.4.4 Cross-Dataset Generalization

In real clinical practice, a model is unlikely to encounter images originating from a single scanner or medical institution. Therefore, beyond evaluating in-domain performance, it is essential to assess how well the model generalizes across different imaging sources. It is very important to check how well the model works when the data distribution changes [5]. Different centers use different hardware (film vs.

digital), acquisition protocols, and they also serve different patient groups. Hence there is always some domain gap between datasets such as DDSM, INbreast, and MIAS [88].

A model that performs well on its training dataset but fails when deployed in a new hospital. To study this, we performed a train-on-two, test-on-the-third evaluation. In each experiment, MsFuNet and the baselines were trained on two datasets together and then tested on the remaining unseen dataset. This setup is closer to a realistic deployment scenario, where a model is trained on data from two centers and then applied at a third center for the first time.

(a) Segmentation (Cross-Dataset Performance): For segmentation, we used mean Pixel Accuracy (mPA), mean Intersection over Union (mIoU), and mean Dice Coefficient (mDice) for over all test cases and the results are shown in Table 4.18. Formally, given prediction masks $P$and ground-truth masks G, these metrics are defined in

$$mPA = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \quad (4.53)–$$

where i indexes test cases.

- **Mean IoU:**

$$mIoU = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i + FN_i} \quad (4.54)$$

- **Mean Dice:**

$$mDice = \frac{1}{N}\sum_{i=1}^{N}\frac{2TP_i}{2TP_i + FP_i + FN_i} \quad (4.55)$$

Table 4.18: Cross-dataset segmentation performance (mPA, mIoU, mDice)

Train Set Combination	Test Set	Model	mPA (%)	mIoU	mDice
DDSM + INbreast	MIAS	U-Net	78.5	0.73	0.75
		MammoUNet	84.5	0.79	0.81
		MsFuNet	**90.5**	**0.84**	**0.85**
DDSM + MIAS	INbreast	U-Net	75.2	0.68	0.7
		MammoUNet	81.0	0.75	0.77

		MsFuNet	**87.1**	**0.81**	**0.82**
INbreast + MIAS	DDSM	U-Net	71.2	0.66	0.68
		MammoUNet	78.5	0.73	0.74
		MsFuNet	**85.0**	**0.79**	**0.8**

Main observations:

- MsFuNet gives the best mPA and mDice in all test domains, improving over MammoUNet by roughly 5–7% mPA and about 0.05 mDice.
- The hardest case is when INbreast is the test set (trained on DDSM + MIAS). Here, all models drop more, which shows that moving from film-based data to high-resolution FFDM images is quite challenging.
- Standard baselines like U-Net and ResUNet show strong performance drops under domain shift, with mDice often < 0.70, which confirms that they are quite sensitive to acquisition and resolution changes.

(b) Classification (Cross-Dataset Performance): For classification, we measured mean Precision (mPrec), mean Recall (mRec), mean F1-score (mF1), and Accuracy (Acc) across cross-dataset splits and results are presented in Table 4.19. The definitions for mPrec and mRec follow

$$mPrec = \frac{1}{c}\sum_{c=1}^{C}\frac{TP_c}{TP_c+FP_c}, \quad mRec = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c+FN_c} \qquad (4.56)$$

$$mF1 = \frac{2\cdot mPrec\cdot mRec}{mPrec+mRec}, \quad Acc = \frac{\sum_c(TP_c+TN_c)}{\sum_c(TP_c+TN_c+FP_c+FN_c)} \qquad (4.57)$$

where C is the number of classes (binary for DDSM/MIAS and multiclass for INbreast).

Table 4.19: Cross-dataset classification performance (mPrecision, mRecall, mF1, Acc)

Train Set Combination	**Test Set**	**Model**	**mPrec (%)**	**mRec (%)**	**mF1 (%)**	**Acc (%)**
DDSM + INbreast	MIAS	ResNet-50	86.3	83	84.6	84.1
		Swin Transformer	91.5	87.1	89.2	92.2
		MsFuNet	**96**	**94.6**	**95.3**	**94.5**

DDSM + MIAS	INbreast	ResNet-50	82.1	81	81.5	81.4
		Swin Transformer	89.9	85.7	87.7	90.1
		MsFuNet	**92.1**	**91.4**	**91.8**	**91.6**
INbreast + MIAS	DDSM	ResNet-50	79.2	80.3	79.7	79.5
		Swin Transformer	87.8	84.4	86.1	85.7
		MsFuNet	**90.5**	**89.8**	**90.1**	**89.9**

Main observations:

- Across all cross-dataset tests, MsFuNet achieves mF1 > 90%, clearly higher than the strong baselines ResNet-50 and Swin Transformer.
- When MIAS is the test set, most models suffer because of its lower resolution and different appearance, but MsFuNet still keeps high recall (94.6%), which means it remains sensitive to malignant cases even under strong domain shift.
- For the multiclass INbreast task, MsFuNet reaches macro-F1 ≈ 91.8%, which shows that it treats normal, benign, and malignant classes in a balanced way, instead of focusing only on the majority class.

(c) Sensitivity to Dataset Shifts: From these experiments, three main conclusions drawn are:

1. When the model is trained exclusively on film-based images (DDSM, MIAS) and subsequently evaluated on high-resolution FFDM data (INbreast), its performance shows the greatest decline. This highlights that the importance of incorporating the high-quality digital mammography data during the training when the intended deployment setting involves FFDM systems.
2. Despite the low resolution and limited number of samples in MIAS dataset, MsFuNet maintains the strong segmentation and classification performance when MIAS dataset serves as the test domain. This behaviour demonstrates that the Domain-Agnostic Regularization Strategy effectively stabilizes performance under significant variations in image appearance.
3. By combining global tissue structures with local lesion details, MsFuNet reduces both false positives and false negatives across domains [93]. In

contrast, ResNet-50 frequently fails to detect certain malignant regions, resulting in elevated false negatives, whereas Swin Transformer is prone to overreacting to noise, leading to increased false positives.

4.4.5 Qualitative Analysis

Even though numerical metrics are important, they do not always show how the masks look or whether the model focuses on clinically meaningful regions. For that reason, we carried out a qualitative analysis of MsFuNet's outputs. We compared MsFuNet with ground truth (GT) and with some of the strongest baselines (MammoUNet, D-UNet, Swin Transformer, etc.) on representative cases from DDSM, INbreast, and MIAS. The selected examples cover three difficulty levels:

- **Easy** – large, well-defined lesions.
- **Medium** – partially overlapped tumours and mixed tissue.
- **Hard** – scattered micro-calcifications, dense tissue, and very irregular margins.

(a) Qualitative Segmentation on DDSM

As shown in Figure 4.10, in easy cases (large, isolated masses), all models roughly outline the lesion. However, MsFuNet's boundaries are smoother and more faithful to the GT, while MammoUNet sometimes extends the mask slightly beyond the actual border [94].

In medium cases, where lesions partly overlap with surrounding tissue, MsFuNet still preserves a clear boundary. In contrast, D-UNet often shows "leakage" into nearby fibroglandular regions. In hard cases with scattered lesions in dense breasts, MsFuNet performs clearly better. It detects most of the small components, while MammoUNet tends to merge close micro-calcifications into a single blob, and U-Net misses several tiny fragments entirely.

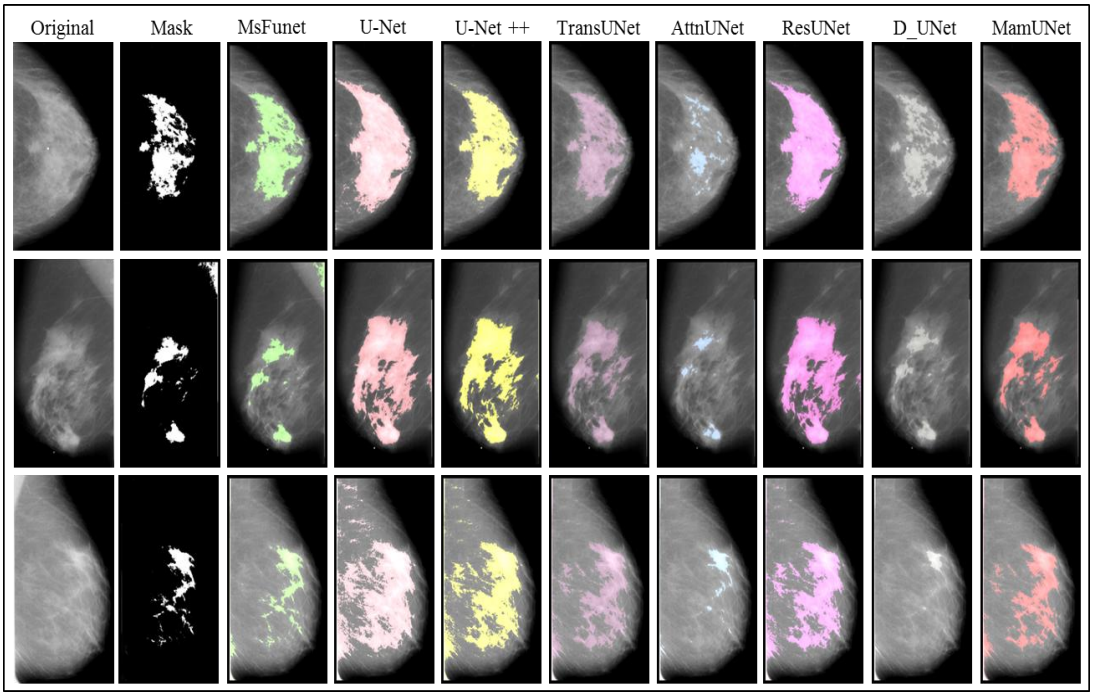

Figure 4.10: DDSM qualitative segmentation grid (GT vs. MsFuNet vs. baselines for easy, medium, and hard cases).

(b) Qualitative Segmentation on INbreast

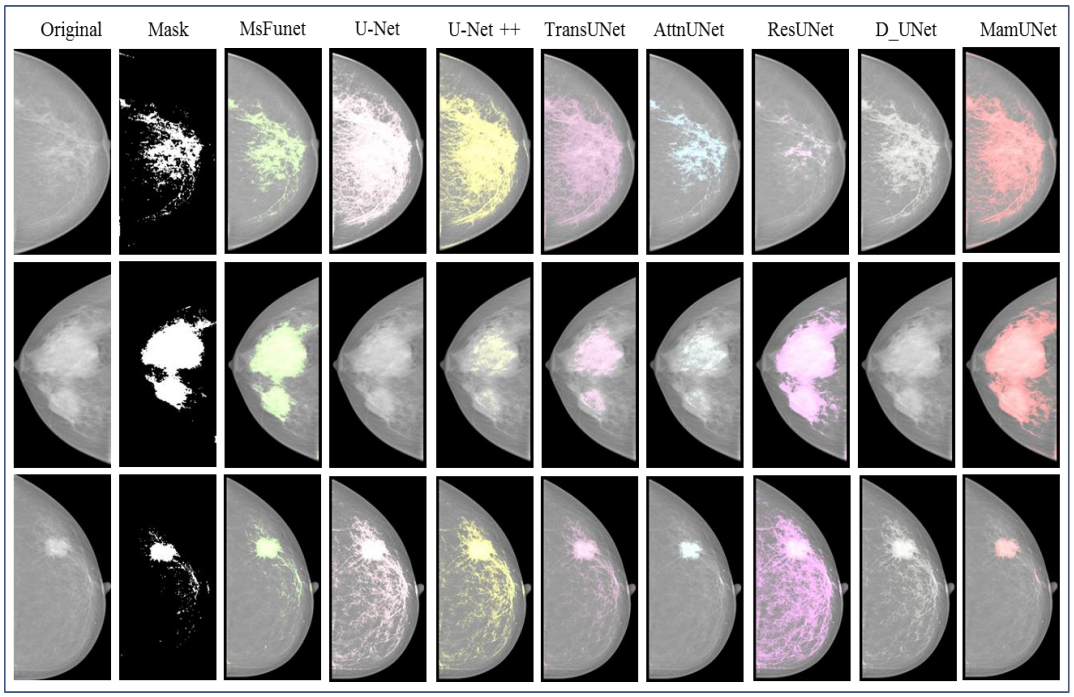

Figure 4.11: INbreast qualitative segmentation grid (GT vs. MsFuNet vs. baselines across different lesion complexities).

On the INbreast dataset (Figure 4.11), which contains high-resolution full-field digital mammography (FFDM) images, MsFuNet produces contours that closely follow the ground truth (GT):

- **Well-defined masses:** MsFuNet generates tight and clean boundaries, whereas MammoUNet's masks appear less precise and are occasionally over-smoothed.
- **Architectural distortions:** MsFuNet effectively highlights subtle linear patterns and star-shaped structures that U-Net often fails to capture.
- **Micro-calcification clusters:** MsFuNet localizes small, bright spots more accurately due to its context-aware feature pyramid. In comparison, D-UNet tends to blur these regions, and ResUNet produces fragmented or incomplete clusters [67].

(c) Qualitative Segmentation on MIAS

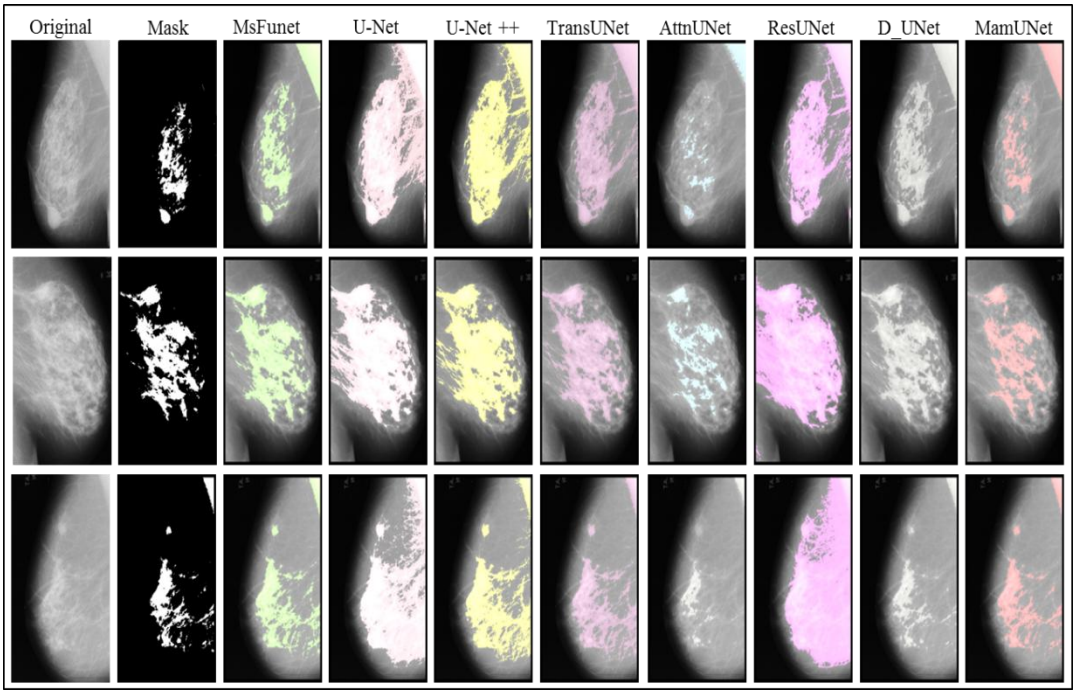

Figure 4.12: MIAS qualitative segmentation grid (GT vs. MsFuNet vs. baseline models under low-resolution conditions).

In the MIAS dataset (Figure 4.12), which consists of lower-resolution images and a smaller sample size the results are described as follows:

- **Easy lesions:** MsFuNet maintains contours closely aligned with GT, while U-Net frequently incorporates additional background around the mass.
- **Medium-density cases:** MsFuNet preserves reasonable locality, whereas MammoUNet sometimes over segments fibroglandular tissue, introducing artifactual extensions.
- **Hard cases (small and fuzzy lesions):** MsFuNet retains good sensitivity, whereas baseline models either miss these subtle regions or merge them with nearby noise, demonstrating their limitations under data-scarce conditions.

(d) Summary of Qualitative Findings

From the visual inspection across the datasets (Figure 4.10, 4.11 and 4.12), a few key points are clear:

- **Boundary precision:** the MsFuNet produces the contours that are generally very close to GT, especially for the irregular or partially hidden lesions.
- **Robustness in dense tissue:** Even under high breast density, MsFuNet keeps the lesion structure more intact. Baseline models are showing more over- or under-segmentation [68].
- **Sensitivity to small lesions:** Scattered micro-calcifications are captured form more reliably by MsFuNet, which supports the claim that its multi-scale and attention mechanisms are working as the intended.
- **Interpretability of classification:** The attention maps are given radiologically meaningful evidence by highlighting the same regions that human readers are would be consider suspicious.

In total, these qualitative results support the quantitative metrics and suggest that the MsFuNet delivers the both strong numerical [60].

4.4.6 Ablation Studies

To quantify the contribution of each major component to the overall performance, ablation experiments are conducted [70]. In each variant, a single module of the MsFuNet was removed, and the resulting model it was retrained under identical settings. The impact of each removal was then assessed by measuring the decrease in

Dice coefficient for the segmentation and F1-score for classification, as well as the recording any change in inference time per image.

Ablation Study

The three ablation variants of the MsFuNet were defined and the results were presented in Table 4.20 as follows:

1. **Without CFPN:** The Context-Aware Feature Pyramid Networks was removed and replaced with the simple encoder–decoder backbones, the resulting in the loss of multiscale feature aggregation.
2. **Without Cross-Attention Fusion:** The cross-attention mechanism block was removed, and global and local features were concatenated by directly [58].
3. **Without Domain Regularization:** The domain-adversarial loss and dataset-specific batch normalization were disabled, leaving the only task-specific objectives.

Table 4.20: Ablation results – changes in average Dice, F1-score, and inference time per image.

Variant	ΔDice (avg)	ΔF1 (avg)	ΔCompute (ms/img)	Notes
w/o CFPN	−0.041	−0.036	−8	Loss of sensitivity to small lesions
w/o Cross-Attention	−0.034	−0.029	−5	Weaker discrimination in dense tissue
w/o Domain Reg.	−0.027	−0.024	−2	Higher sensitivity to dataset shift
Full MsFuNet	—	—	42	Baseline (best overall performance)

Interpretation:

- **CFPN removal:** Eliminating the CFPN caused by the largest Dice drop (~4.1%), confirming that multiscale context is crucial for the accurately delineating tumour boundaries, particularly for the small or subtle lesions.

- **Cross-Attention Fusion removal:** Excluding these modules are reduced the average F1-score by approximately 3%, with the most pronounced effect on the malignant-versus-benign discrimination in dense tissue. Without cross-attention mechanism, the models are struggles to integrate the local details with the global structure.
- **Domain Regularization removal:** Disabling the domain regularization primarily affected by cross-dataset performance. In-domain segmentation and classification remained adequate, but performance under the domain shift deteriorated. This is highlighting the importance of adversarial regularization for handling the dataset variability.

Finally, the ablation study demonstrates that the MsFuNet's strong performance does not rely on any single component. Rather, than the observed gains in both segmentation and classification, particularly under cross-domain conditions, are arise from the synergistic effect of:

- The **Context-Aware CFPN**, which provides rich multiscale features,
- The **Cross-Attention Fusion**, which links global and local cues, and
- The **Domain-Agnostic Regularization**, which ensures robustness across datasets.

Together, these components enable MsFuNet to achieve accurate and generalizable performance across diverse mammography datasets.

4.4.7 Efficiency and Computational Complexity

For real clinical use, a model must be accurate but also computationally reasonable. If a system is too heavy, it will be hard to integrate into routine hospital workflows, especially in busy screening centers. Therefore, we also evaluated the MsFuNet in terms of: number of parameters, FLOPs (floating-point operations), GPU memory usage, throughput (images processed per second), and latency (time per image). These values were compared with strong baselines such as U-Net, TransUNet, and MammoUNet are presented in Table 4.21.

Table 4.21: Computational profile of MsFuNet vs. baseline models

Model	Params (M)	FLOPs (G)	Memory (GB)	Throughput (img/s)	Latency (ms/img)
U-Net	7.8	25.2	3.1	120	8.3
TransUNet	105	170.4	14.6	32	31.2
MammoUNet	9.5	31.8	3.6	98	10.2
MsFuNet	28.6	56.4	5.8	74	13.5

From Table 4.21, MsFuNet sits in a middle region. It is clearly heavier than U-Net and MammoUNet, but still much lighter than transformer-based architectures such as TransUNet, which uses more than 100M parameters and very high FLOPs. With around 28M parameters and 56 GFLOPs, MsFuNet keeps a reasonable trade-off between complexity and performance. It can process about 74 images per second, with a latency of roughly 13–14 ms per image, which is acceptable for batch or near-real-time screening settings [95].

4.4.8 Statistical Significance

To ensure that the observed improvements of MsFuNet over baseline models are not due to random fluctuations, a formal statistical significance analysis was conducted. Comparisons were always made against the best-performing baseline on each dataset.

(a) Paired Tests: For each dataset and evaluation metric, per-fold scores from the 3-fold cross-validation were collected. The following statistical tests were applied:

- **Paired t-test:** Used when the differences in scores approximately followed a normal distribution.
- **Wilcoxon signed-rank test:** Applied when the normality assumption was uncertain.

The paired t-test statistic is defined by incorporating the mean difference across folds, the standard deviation of the differences, and the number of folds.

$$t = \frac{\dot{X}_1 - \dot{X}_2}{\frac{s_d}{\sqrt{n}}} \quad . (4.58),$$

Across all datasets, MsFuNet achieved significantly higher Dice coefficients (segmentation) and F1-scores (classification) than the best baseline, with $p < 0.01$in

nearly all cases, indicating a very low probability that the observed gains are due to chance.

(b) Confidence Intervals: To assess the uncertainty of the performance is estimates, bootstrapping with 1,000 resamples was employed to the compute 95% confidence intervals. The Representative results include:

- **Segmentation (Dice):**
    - MsFuNet: [0.872, 0.895]
    - MammoUNet: [0.843, 0.861]
- **Classification (F1-score):**
    - MsFuNet: [0.930, 0.945]
    - Swin Transformer: [0.895, 0.912]

These intervals are demonstrated that the even when accounting for data variability, MsFuNet consistently outperforms the competing models, are within the minimal overlap between the confidence bands.

(c) Effect Sizes

In addition to p-values, Cohen's d was calculated to quantify the practical significance of the improvements. The formula is provided by

$$d = \frac{\mu_{MsFuNet} - \mu_{Baseline}}{s_p},\ s_p = \sqrt{\frac{s_1^2 + s_2^2}{2}} \quad (4.59),$$

It is combining the mean difference and pooled standard deviation of the two models.

- **Segmentation (Dice):** Cohen's d ranged from the approximately 0.65 to 0.80, considered a large effect.
- **Classification (F1-score):** Cohen's d ranged from 0.72 to 0.84, also indicating the large and practically meaningful of improvement.

Significance Analysis

- MsFuNet's gains over the top baselines models are statistically significant (mostly $p < 0.01$).

- Confidence intervals reveal a clear separation between MsFuNet and the best baselines for both Dice and F1 metrics.
- Effect sizes in the medium-to-large range confirm that the improvements are both statistically and clinically meaningful.
- Additionally, the robustness checks, including the alternative fold splits and removal of outlier cases, did not alter the main conclusion: MsFuNet consistently outperforms baseline models are across the datasets and metrics.

4.5 Chapter Summary

Core framework: This chapter presents MsFuNet, a unified deep learning model that performs both tumour segmentation and malignancy classification on mammograms. Instead of having separate networks for each task, MsFuNet uses:

- a context-aware feature pyramid network (CFPN) backbone to learn multi-scale features,
- a cross-attention fusion module to merge local and global information across different scales.

A shared multi-scale encoder feeds into two task-specific heads:

- one head for pixel-wise tumour segmentation,
- one head for image-level benign/malignant classification.

This multitask learning encourages consistent predictions (the segmentation and classification support each other) and makes more efficient use of features. Additionally, MsFuNet applies domain-agnostic regularization and strong preprocessing (normalization, artifact removal) to improve robustness when moving between datasets.

Datasets and results: MsFuNet is tested on three mammography datasets—DDSM, INbreast, and MIAS—covering diverse imaging conditions and populations.

- For segmentation, it reaches Dice scores around 0.89–0.90, slightly but consistently better than advanced CNN and transformer baselines.

- For malignancy classification, MsFuNet achieves an F1-score of about 93.7%, compared with roughly 90% for the next best model, and maintains a low false-negative rate, which is very important for screening.

In cross-dataset experiments, MsFuNet keeps high accuracy and recall where many conventional models suffer large performance drops under domain shift. Chapter 4 shows that multi-scale feature fusion plus joint segmentation–classification training leads to more robust, context-aware predictions and better generalization.

CHAPTER – 5 A NOVEL DEEP LEARNING ARCHITECTURE FOR BREAST CANCER SEGMENTATION AND CLASSIFICATION USING SPATIAL AND TEMPORAL FEATURES

5.1 Introduction

5.1.1 Background and Clinical Motivation

Breast cancer remains one of the most significant global health challenges and continues to exert the substantial impact on morbidity, mortality, and the overall the quality of life of women [96]. These assessments demonstrate that the breast cancer is not only that the most frequently diagnosed cancer among women but also one of the leading causes of cancer-related to deaths [97]. Early and accurate diagnosis is essential for the improving patient survival and enabling more effective treatment planning. In routine clinical practice, mammography remains the gold standard for screening and diagnostic. Nevertheless, the interpretation of the mammograms can be highly challenging. Overlapping the breast tissues, poorly defined tumour margins, and variations in breast density can complicate image evaluation, even for the experienced radiologists.

In this context, the segmentation and classification of breast lesions constitute critical components of computer-aided diagnosis systems designed to the support radiologists in clinical decision-making. The Segmentation aims to outline the lesion boundary precisely, whereas the classification decides whether the abnormality is benign or malignant. These two tasks are support each other and together provide the basis for reliable decision support in oncology.

With the progress in deep learning, especially convolutional neural networks (CNNs), CAD systems are have improved a lot. Most current approaches, however, the mainly focus on spatial information inside a single image. They usually ignore the temporal relationships, for example, across different mammographic views or across feature maps at different levels inside the network. A joint analysis of spatial and temporal aspects can give a more complete description of tumour behaviour and can improve both segmentation and classification results [55].

5.1.2 Problem Statement and Gaps

Even though CNN-based and other deep learning models have made strong progress, several important gaps still remain in breast cancer image analysis:

- **Limited temporal modelling:** Classical CNNs are mainly designed for static images. They are not well-suited to capture the sequential or temporal dependencies, which are often useful when we compare multiple mammographic views or analyses feature sequences. As a result, these models may miss evolving contextual patterns that could help to separate benign from malignant lesions.
- **Strong focus on spatial features only:** Most current segmentation and classification frameworks rely heavily on spatial cues such as shape, texture, and intensity. These cues are important, but in many borderline cases they are not enough. Subtle differences between lesion types may only become clear when temporal or sequential information is also considered.
- **Overlapping feature representations:** Benign and malignant lesions can share similar spatial characteristics. When the model has access only to static spatial features, their representations may overlap in feature space, which leads to frequent misclassifications. Temporal cues could help to break this overlap, but they are usually not exploited.
- **Limited interpretability and clinical trust:** Many deep learning models act as "black boxes". Without tools such as attention maps or other visual explanations, clinicians may find it hard to trust or understand the model's decisions. This lack of interpretability makes it difficult to integrate such systems into routine practice.

These issues summarized that purely spatial deep learning pipelines are not sufficient. There is a clear need for architectures that can jointly use spatial patterns and temporal relationships. So that both segmentation and classification become more accurate and more clinically reliable.

5.1.3 Objectives and Contributions

To address the above challenges, this thesis proposes a new framework called Spatial–Excitation Convolutional Recurrent Neural Network (SE-CRNN). The main

idea is to combine spatial attention with temporal modelling in a unified design. SE-CRNN is built to overcome the typical limits of standard CNN-based systems and to provide outputs that are more robust and more useful for clinical decision-making (Table 5.1). The main objectives and contributions are listed below as:

1) Enhancing Spatial Features with SE Blocks: SE-CRNN uses Squeeze-and-Excitation (SE) blocks to adaptively adjust the importance of each feature channel. In simple terms, the network learns to highlight the most informative tumour-related channels and to suppress background or noisy channels.

- This channel-wise reweighting helps the model focus on tumour regions instead of irrelevant tissue.
- As a result, lesion segmentation becomes more precise and the number of false positives can be reduced.

2) Modelling Temporal Dependencies with Bidirectional RNNs: To capture temporal or sequential information, SE-CRNN integrates Bidirectional Recurrent Neural Networks (BRNNs).

- BRNNs process information in both forward and backward directions, so they can use context from earlier and later steps in the sequence.
- In our case, this allows the model to combine information across multiple mammographic views or along sequential feature maps, giving a richer understanding of lesion characteristics [98].
- This temporal context supports more confident and stable classification decisions.

3) Task-Specific Architectures for Segmentation and Classification: SE-CRNN is organized into two main task-oriented branches:

- **SUNet (Segmentation U-Net with SE blocks)**
    - A U-Net based architecture enhanced with SE modules.
    - Designed specifically for accurate segmentation of tumour boundaries.

- The SE blocks help SUNet focus on lesion areas and maintain sharper and more reliable contours.

- **TeResNet (Temporal ResNet with BRNN integration)**
  - A classification framework that combines ResNet's strong spatial feature extraction with BRNN-based temporal learning.
  - Aims to improve benign vs malignant discrimination by using both spatial appearance and temporal/sequence information.

Together, SUNet and TeResNet form the full SE-CRNN framework for joint segmentation and classification.

4) Comprehensive Evaluation on Multiple Datasets: To check the generality and robustness of SE-CRNN, extensive experiments are carried out on: CBIS-DDSM, INbreast, and combined or integrated datasets. Performance is compared with several state-of-the-art methods using common metrics such as:

- Dice Similarity Coefficient (DSC) and Intersection over Union (IoU) for segmentation,
- Accuracy, Sensitivity, and Specificity for classification.

Table 5.1: Overview of Objectives and Contributions of SE-CRNN

Contribution Area	Proposed Approach	Expected Impact
Spatial feature enhancement	Squeeze-and-Excitation (SE) blocks	Better focus on tumour regions; improved segmentation
Temporal modelling	Bidirectional Recurrent Neural Networks	Capture sequential dependencies; more robust classification
Segmentation framework	SUNet (SE-augmented U-Net)	Higher accuracy in delineating benign/malignant lesion edges
Classification framework	TeResNet (ResNet + BRNN integration)	Reliable benign vs malignant discrimination
Dataset validation	CBIS-DDSM, INbreast, and integrated datasets	More generalizable and clinically adaptable CAD system

The roadmap of this chapter starts from the clinical background and problem definition, then introduces the SE-CRNN design, followed by details of the

experimental setup, quantitative and qualitative results, and a discussion of clinical implications. This structure is intended to present the proposed framework in a clear and logical way within the overall thesis.

5.2 SE-CRNN Methodology

5.2.1 Architectural Overview

The Spatial–Excitation Convolutional Recurrent Neural Network (SE-CRNN) is designed as a single framework (Figure 5.1) that handles both tumour segmentation and classification in breast cancer imaging [99]. The key idea is to join spatial attention and temporal modelling, so that the system is not limited to only static spatial patterns like a traditional CNN.

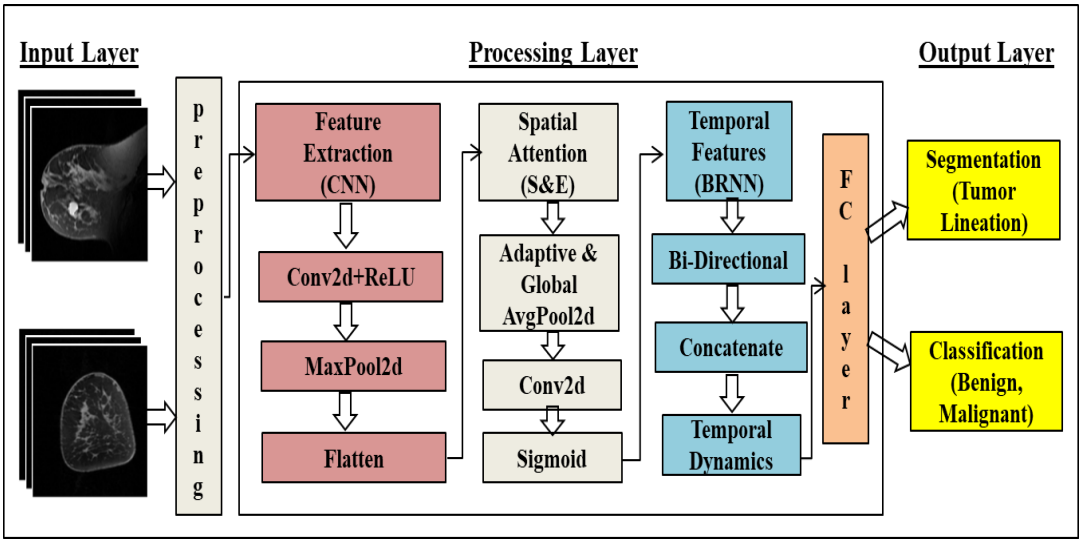

Figure 5.1: Spatial and Temporal Features-Based Breast Cancer Segmentation and Classification Architecture

At a high level, the SE-CRNN pipeline has the following main stages:

1. **Preprocessing and Normalization:** Input mammograms are first standardized in terms of size, resolution, and contrast. This step makes images from different datasets more consistent, so that the later feature extraction and sequence modelling become more stable.

2. **CNN Backbone with SE Blocks:** A convolutional backbone extracts hierarchical spatial features. Inside this backbone, Squeeze-and-Excitation (SE) blocks are inserted to perform channel-wise attention. In practice, this means the network learns to give higher weight to channels that carry

tumour-relevant information, and lower weight to channels dominated by background tissue.

- **Bidirectional Recurrent Neural Network (BRNN):** The refined feature maps are then reshaped into a sequence and passed through a BRNN. This stage captures temporal or sequential dependencies, between multiple mammographic views (e.g., CC and MLO), or across feature hierarchies.

3. **Dual Task-Specific Heads**

- **SUNet (Segmentation U-Net with SE integration):** Uses the enriched spatial-temporal features to produce binary tumour masks, focusing on accurate lesion localization.
- **TeResNet (Temporal ResNet):** Uses temporally enhanced features to classify tumours as benign or malignant, aiming for higher confidence and reliability.

5.2.2 Preprocessing and Normalization

Preprocessing is a very important step because mammograms usually come from different scanners, different protocols, and different contrast settings. If we feed them directly into a deep model, this variability can hurt in both training and inference. As shown in figure 5.2, SE-CRNN uses a simple but effective preprocessing pipeline with two main parts:

1. Image Resizing with Bicubic Interpolation
2. Normalization using Adaptive Histogram Equalization (AHE)

(a) Image Resizing with Bicubic Interpolation: Raw mammograms often have different original sizes and aspect ratios. Training a network directly on such images would be inefficient and could also make the learned filters inconsistent. To avoid this, all images are resized to a fixed resolution, for example 256 × 256 pixels, using bicubic interpolation. This gives smoother and more natural results than simple nearest-neighbor or bilinear interpolation. The general form of the interpolation can be written as

$$I'(x,y) = \sum_{i=-1}^{2}\sum_{j=-1}^{2} w(i,j) \cdot I(x+i, y+j) \quad (5.1),$$

where the original pixel intensities are combined, the interpolation weights depend on cubic polynomials in the horizontal and vertical directions, and the final resized intensity is obtained as a weighted sum. This step standardizes the spatial size across datasets but still preserves the basic tumour morphology, which is important for segmentation.

(b) Normalization with Adaptive Histogram Equalization (AHE): Even after resizing, many mammograms suffer from low or uneven contrast. Tumour regions can blend into dense tissue, making boundaries very hard to see. To improve this, Adaptive Histogram Equalization (AHE) is applied. Unlike global histogram equalization (which adjusts contrast over the whole image in one step), AHE works on small local regions, usually called tiles or blocks. For each local block, a separate histogram is computed and equalized, so that local details become more visible. The transformation can be written in

$$I''(x,y) = \frac{CDF\left(I'(x,y)\right) - CDF_{min}}{N - CDF_{min}} \times (L-1) \qquad (5.2),$$

Where the mapping is based on the local cumulative distribution function, N is the number of pixels in the local region, and L is the total number of grey levels. In SE-CRNN, AHE is applied with, region size of 8×8blocks, clip limit of around 2.0 (Table 5.2) to avoid over-amplifying noise. The raw mammogram, the resized version, and the final AHE-normalized output.

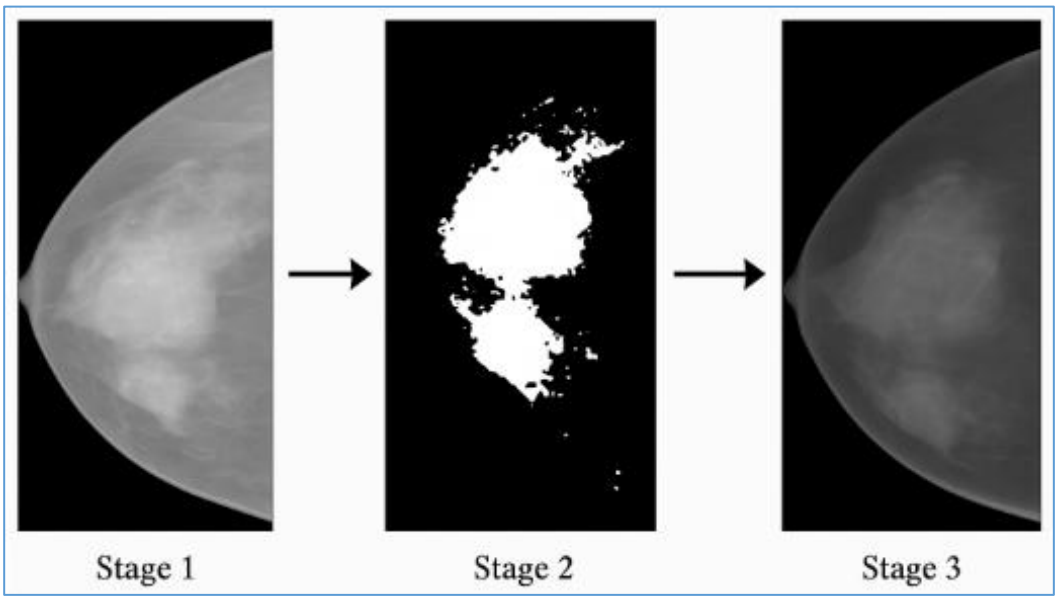

Figure 5.2: Preprocessing workflow with sample images shows three stages

Table 5.2: Preprocessing parameters used in SE-CRNN

Step	Technique	Parameters	Purpose
Image Resizing	Bicubic interpolation	Final size: 256 × 256	Standardize image size; keep morphology
Normalization	Adaptive Hist. Equal.	Tile size: 8 × 8; Clip limit: 2.0	Improve local contrast; emphasize lesions

Together, bicubic resizing and AHE ensure that the mammograms entering SE-CRNN are size-normalized and contrast-enhanced, which makes the later spatial–temporal modelling more effective.

5.2.3 CNN Feature Extraction

After preprocessing, the first major learning stage is featuring extraction with a CNN [100]. CNNs are a natural choice here because they can automatically learn hierarchical representations:

- **low-level:** edges, simple textures, bright spots,
- **mid-level:** shapes of masses, local patterns,
- **high-level:** more abstract tumour characteristics.

This is especially important in mammography, where fine details such as spiculated margins, irregular contours, and subtle density changes are crucial to distinguish benign from malignant lesions. The basic convolution operation for layer $l$at spatial location (i, j) can be written as

$$F_l(i,j) = \sigma(\sum_{m=1}^{M}\sum_{n=1}^{N} W_l(m,n) \cdot X(i+m, j+n) + b_l) \qquad (5.3),$$

Where the filter weights slide over the input feature map, a bias term is added, and a non-linear activation function $\phi(\cdot)$ (for example ReLU) is applied. To gradually reduce the spatial resolution and make the representation more compact and translationally robust, pooling layers are applied. For max pooling, the output is given by

$$P(i,j) = max_{(u,v)\epsilon R} F(i+u.j+v) \text{ (5.4)},$$

where the maximum value inside a region R is chosen. This helps the network become less sensitive to small spatial shifts. In SE-CRNN, two main activation functions are used:

- **ReLU** – applied after convolution layers to introduce non-linearity and help avoid vanishing gradients;

$$ReLU(x) = max(0, x), \quad (5.5)$$

- **Sigmoid** – used in later stages for probability mapping (e.g., mask generation or final class scores).

$$\sigma(x) = \frac{1}{1+e^{-x}} \quad (5.6)$$

After several convolution + pooling stages, the CNN produces a set of multi-scale feature maps. These form the input to the SE blocks (for spatial attention) and then to the BRNN (for temporal modelling).

5.2.4 Spatial Attention via SE Block

Although the CNN learns a rich representation, not all feature channels are equally informative for tumour detection [101]. Some channels mainly capture background tissue, while others strongly respond to suspicious masses or calcifications. To handle this, SE-CRNN integrates SE blocks inside the backbone. As shown in figure 5.3, the SE block applies channel-wise attention, so that the network can adaptively boost useful channels and suppress less relevant ones [102].

The SE block has two main steps:

1. **Squeeze – Global Information Embedding:** For each channel c, a global average pooling is computed across spatial dimensions, producing a single scalar descriptor:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_c\,(i, j) \quad (5.7)$$

where X_c is the c-th channel of the feature map and its height and width are averaged. This gives a compact summary of the global response of that channel.

2. **Excitation – Adaptive Recalibration:** The vector of channel descriptors is passed through two fully connected layers with non-linear activations:

$$s = \sigma(W_2\ \delta(W_1 z)) \quad (5.8)$$

where W_1 and W_2 are weight matrices, $\delta(\cdot)$ is the ReLU, and $\sigma(\cdot)$ is the sigmoid. The output s contains scaling factors for each channel. Finally, each channel is rescaled as:

$$\tilde{X}_c = s_c \cdot X_c \qquad (5.9)$$

This means channels with high s_c are strengthened, while channels with low s_c are suppressed. In mammograms, this process tends to emphasize subtle lesion areas and reduce background clutter.

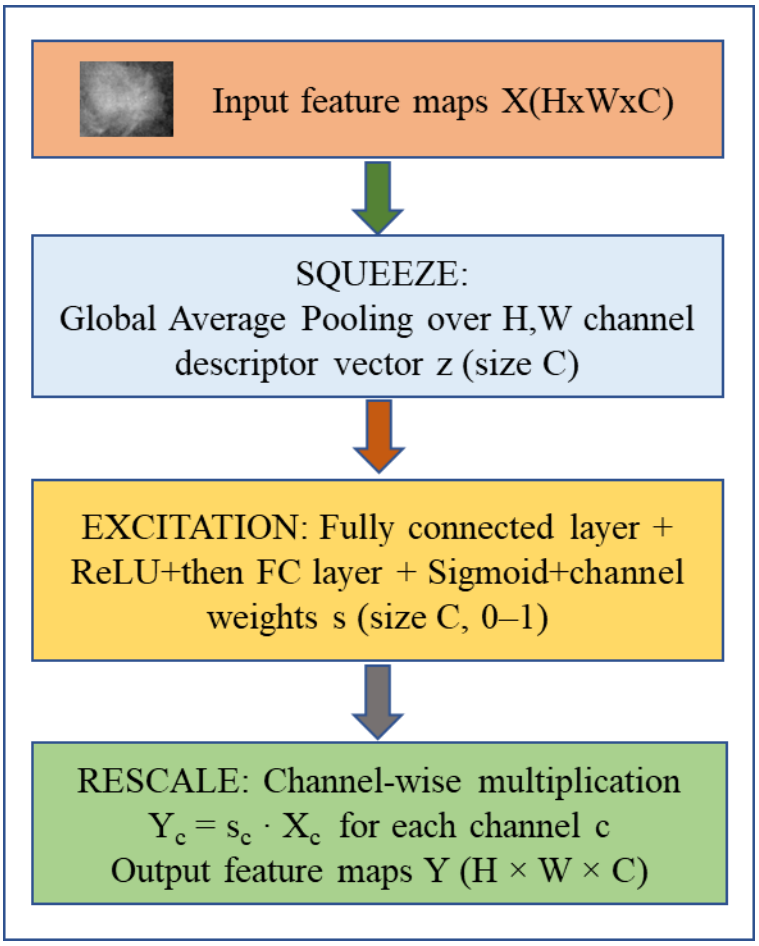

Figure 5.3: SE block structure (squeeze → excitation → rescale)

5.2.5 Temporal Feature Extraction with BRNN

Although mammograms are static images, the feature maps extracted by convolutional layers can be represented as ordered sequences. Bidirectional Recurrent Neural Networks (BRNNs) are employed to model the dependencies among these sequential feature representations, enabling the capture of long-range contextual information that may not be adequately learned by convolutional operations alone. The term temporal features refers to the sequential relationships

among spatial feature vectors, multi-scale representations, or multi-view mammographic features rather than actual temporal changes. By processing feature sequences in both forward and backward directions, the BRNN enhances contextual understanding and improves classification performance, particularly for subtle and complex breast lesions.

Even after spatial attention, the features are still mostly static. However, mammography naturally contains sequential relationships, for example:

- paired CC and MLO views of the same breast,
- feature hierarchies across layers,
- or time-ordered examinations in longitudinal follow-up.

To capture such dependencies, SE-CRNN uses Bidirectional Recurrent Neural Networks (BRNNs) on top of the SE-refined feature maps. First, the 2D feature maps are reshaped into a sequence:

$$\{x_t\}_{t=1}^{T} \qquad (5.10)$$

where each x_t is a feature vector at timestep t, and T is the sequence length (e.g., over views or spatial positions). The BRNN has two separate recurrent units:

- a forward unit that processes from $t = 1 \rightarrow T$,
- a backward unit that processes from $t = T \rightarrow 1$.

Forward pass: The forwards pass is defined as:

$$h_t^f = \emptyset\left(W_{xh}^f x_t + W_{hh}^f h_{t-1} + b^f\right) \qquad (5.11)$$

where h_t^f is the forward hidden state, $W_{xh}^f x_t$ and $W_{hh}^f h_{t-1}$ are input and recurrent weights, and $\emptyset$ is the activation function (tanh).

Backward pass: Similarly, the backwards pass is defined as:

$$h_t^b = \emptyset\left(W_{xh}^b x_t + W_{hh}^b h_{t-1} + b^b\right) \quad (5.12)$$

where h_t^b captures dependencies from the future context. The final output at timestep t is the concatenation of forward and backward states:

$$h_t = \left[h_t^f : h_t^b\right] \qquad (5.13)$$

This representation contains information from past and future within the sequence, giving a more complete temporal context. In the context of SE-CRNN, the BRNN helps to:

- relate the patterns between different views of the same breast,
- capture long-range dependencies across feature slices, and
- Distinguish subtle benign structures from malignant progression in a better way.

The BRNN temporal modelling, converts the spatially attended features into sequences and processed in both directions. They contain a forward-only RNN with a bidirectional RNN, highlighting that the bidirectional version captures more complete information for lesion classification.

Table 5.3: Summary of Feature Extraction and Attention Modules in SE-CRNN

Module	**Function**	**Key Output**
CNN (Conv + Pooling)	Learn local and hierarchical spatial features	Multi-level feature maps
Activation Functions (ReLU, Sigmoid)	Introduce non-linearity; map to probabilities	Activated feature maps and final probabilities
SE Block (Squeeze + Excitation)	Channel-wise recalibration (spatial attention)	Tumour-focused, refined feature maps
BRNN (Forward + Backward)	Temporal sequence modelling	Contextualized spatio-temporal representations

As shown in table 5.3, the CNN, SE block, and BRNN form the core backbone of SE-CRNN. They ensure that before the features reach the SUNet and TeResNet heads, both fine spatial cues and long-range temporal dependencies are already encoded in a compact and discriminative representation.

5.2.6 Tumour Segmentation with SUNet

Accurate segmentation of suspicious regions in a mammogram is very important, because it gives a clear outline of the tumour and helps in later clinical decisions and treatment planning. In the SE-CRNN framework, this task is handled by SUNet, which is basically a U-Net architecture enhanced with SE blocks. The SE blocks help the network to strengthen channels that are truly related to the lesion and to

suppress channels that mainly describe background tissue, especially in dense and low-contrast breasts. SUNet keeps the classical encoder–decoder structure of U-Net:

- **Encoder path:** a stack of convolution + pooling layers that gradually reduce spatial resolution while extracting higher-level semantic features [103].
- **Decoder path:** transposed convolutions (upsampling) combined with convolution layers to rebuild a full-resolution feature map.
- **Skip connections:** at each stage, encoder features are concatenated with decoder features. This preserves fine details and helps to segment small or irregularly shaped tumours that might otherwise disappear during pooling.

The final output of SUNet is a binary mask, where each pixel is labelled as tumour (1) or non-tumour (0). To train SUNet, we use a hybrid loss that combines Dice loss and cross-entropy loss:

$$\mathcal{L}_{\text{seg}} = \alpha\, \mathcal{L}_{\text{Dice}} + (1-\alpha)\, \mathcal{L}_{\text{CE}} \qquad (5.14)$$

where α is a weighting factor (set to 0.5 in my experiments, so both terms have equal importance). The Dice loss encourages high overlap between the predicted mask $\hat{p}_i$ and the ground-truth mask g_i:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_i \hat{p}_i g_i + \varepsilon}{\sum_i \hat{p}_i + \sum_i g_i + \varepsilon} \qquad (5.15)$$

where ε is a small constant to avoid division by zero. The cross-entropy loss focuses on pixel-wise classification errors:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N} [\, g_i \log(\hat{p}_i) + (1-g_i)\log(1-\hat{p}_i)\,] \qquad (5.16)$$

where N is the number of pixels. In simple terms, the Dice term takes care of global region overlap, while cross-entropy improves local pixel accuracy. Using both in combination helps SUNet to produce masks that are not only accurate in area but also have good boundary quality.

5.2.7 Tumour Classification with TeResNet

After segmentation, SE-CRNN performs tumour classification, deciding whether a given case is benign or malignant. This is done using TeResNet, which is a ResNet-

based classifier augmented with SE blocks and the spatio-temporal features coming from the BRNN. The main steps in TeResNet are:

1. **Feature input:** the refined features from the CNN + SE + BRNN backbone are fed into a ResNet.
2. **Residual learning:** ResNet blocks use skip connections to keep gradient flow stable, allowing deeper architectures without vanishing gradients.
3. **Fully connected head:** after global average pooling, the feature vector is passed through one or more dense layers.
4. **Sigmoid output:** the last layer uses a sigmoid activation to produce a probability $\hat{y}_i \in [0,1]$that the i-th sample is malignant.

The classification branch is trained with binary cross-entropy (BCE):

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N}\sum_{i=1}^{N}[\, y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\,] \qquad (5.17)$$

where $y_i \in \{0,1\}$ is the true label of the i-th sample (0 = benign, 1 = malignant), and $\hat{y}_i$ is the predicted probability of malignancy. This loss forces the model to give high confidence to correct predictions and penalizes wrong ones.

5.2.8 Joint Optimisation and Loss Design

Because SE-CRNN performs two tasks at the same time—segmentation (SUNet) and classification (TeResNet)—the training procedure must balance both objectives fairly. For this reason, we use a joint loss function (Table 5.4):

$$\mathcal{L}_{\text{total}} = \lambda_{\text{seg}}\,\mathcal{L}_{\text{seg}} + \lambda_{\text{cls}}\,\mathcal{L}_{\text{cls}} \qquad (5.18)$$

where $\mathcal{L}_{\text{seg}}$ is the hybrid Dice + cross-entropy loss from.

$$L_{seg} = \alpha L_{Dice} + (1-\alpha)L_{CE}\ , \quad (5.19)$$

where α is a weighting factor (0.5 in our experiments). The Dice loss is given by:

$$L_{Dice} = 1 - \frac{2\sum_{i=1}^{N} p_i g_i}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i} \qquad (5.20)$$

where $p_i g_i$ representing predicted and ground truth labels for pixel i. This term ensures overlap maximization between prediction and ground truth. The cross-entropy loss is defined as:

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}[g_i\, log(p_i) + (1-g_i)\, log\,(1-p_i)] \quad (5.21)$$

which penalizes misclassified pixels at the individual level. $\mathcal{L}_{\text{cls}}$ is the BCE loss from

$$L_{cls} = -\frac{1}{N}\sum_{i=1}^{N}[y_i\, log\dot{y}_i + (1-y_i)\, log(1-\dot{y}_i)] \quad (5.22)$$

where λ_{seg} and λ_{cls} are weights that control the contribution of each term. In my implementation, I set $\lambda_{\text{seg}} = \lambda_{\text{cls}} = 0.5$, so segmentation and classification have equal influence during training. This avoids a situation where the model optimizes only one task and ignores the other.

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{\dot{m}_t}{\sqrt{\dot{v}_t + \epsilon}} \quad (5.23)$$

For optimization, the Adam optimizer is used, since it provides adaptive learning rates and uses momentum on both the first and second moments of the gradients. The parameter updates for θ at step t follows the standard Adam rule:

$$\theta_{t+1} = \theta_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon} \quad (5.24)$$

where η is the learning rate, and $\hat{m}_t$, $\hat{v}_t$ are the bias-corrected estimates of the first and second moments of the gradient. A step-decay learning rate schedule is used: the learning rate is reduced by a factor of 0.5 after every 20 epochs if the validation loss does not improve. In addition, early stopping is applied so that training is stopped when the validation performance plateaus, which helps to avoid overfitting.

Table 5.4 Summarization of the overall loss and optimization strategy

Task	Loss Function(s)	Main Purpose	Optimizer / Settings
Segmentation	Dice + Cross-Entropy (hybrid)	Balance region overlap and pixel-wise accuracy	Adam, LR = 0.001, (\beta_1=0.9, \beta_2=0.999)
Classification	Binary Cross-Entropy (BCE)	Benign vs malignant discrimination	Adam, same settings
Joint Training	Weighted sum (L_{total})	Balance both tasks in a single framework	Step-decay LR; Early stopping

In summary, shows Table 5.4 SUNet and TeResNet, trained together under this joint loss, allow SE-CRNN to deliver precise tumour boundaries and reliable diagnostic labels at the same time [104]. This dual ability is important for real

clinical deployment, where segmentation helps with treatment planning and classification supports the final diagnostic decision.

5.3 Experimental Setup

This section explains the datasets, train–test splits, hardware and software environment, training settings, evaluation metrics, and the baseline methods used to evaluate the proposed SE-CRNN framework (SUNet + TeResNet). The idea is to study both segmentation and classification performance, and also to see how well the model generalizes across different mammography datasets.

5.3.1 Datasets and Splits

(a) Datasets: For experiments, we used three dataset configurations that helps to check both within-dataset performance and cross-dataset robustness (See Figure5.4):

1. **CBIS-DDSM (Curated Breast Imaging Subset of DDSM)**
    - Contains film/converted mammograms.
    - Includes pathology-verified labels and pixel-level mass masks.
    - Suitable for both segmentation and classification tasks.
2. **INbreast**
    - A high-quality full-field digital mammography (FFDM) dataset.
    - Provides expert-annotated masks and rich lesion information.
    - Useful to test performance on modern digital systems.
3. **Integrated (CBIS-DDSM + INbreast)**
    - A merged dataset that combines images from both CBIS-DDSM and INbreast.
    - Designed to test robustness under mixed acquisition protocols, different scanners, and variable imaging conditions.

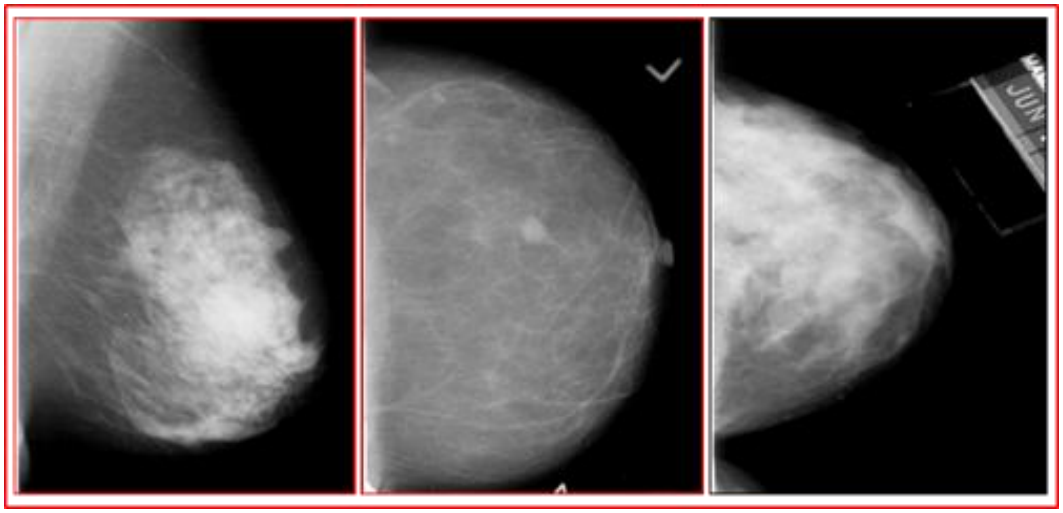

Figure 5.4: Typical mammography image from the CBIS-DDSM

From figure 5.5 these three settings together help to analyses how SE-CRNN behaves when the input data distribution changes between film-based and digital mammograms.

(b) Splits: To keep the experiments reproducible and fair, we follow a consistent splitting strategy for all three configurations:

- **CBIS-DDSM:** (70% training, 15% validation, 15% testing) stratified by benign / malignant labels.
- **INbreast:** (70% training, 15% validation, 15% testing) also stratified to preserve class balance.
- **Integrated set (CBIS-DDSM + INbreast):** Images from both datasets are pooled together with (70% training, 15% validation, 15% testing).

The splits are case-aware and all views from the same exam stay in the same partition (train, validation, or test). This is important because it avoids data leakage; otherwise, views from the same patient might appear in both training and test sets, which would artificially inflate the reported accuracy. The total dataset statistics, including approximate case and image counts, typical resolution, mask availability, and split ratios, are summarized in Table 5.5.

Table 5.5 — Dataset statistics (cases, images, resolution, masks, splits)

Dataset	Cases (≈)	Images	Typical Resolution (px)	Mask Availability	Train / Val / Test
CBIS-DDSM	1,250	1,800	3000 × 2000 → resized	Yes (mass masks)	70 / 15 / 15
INbreast	410	410	3328 × 4084 → resized	Yes (masks)	70 / 15 / 15
Integrated	≈1,660	≈2,210	mixed resolutions → resized	Yes	70 / 15 / 15

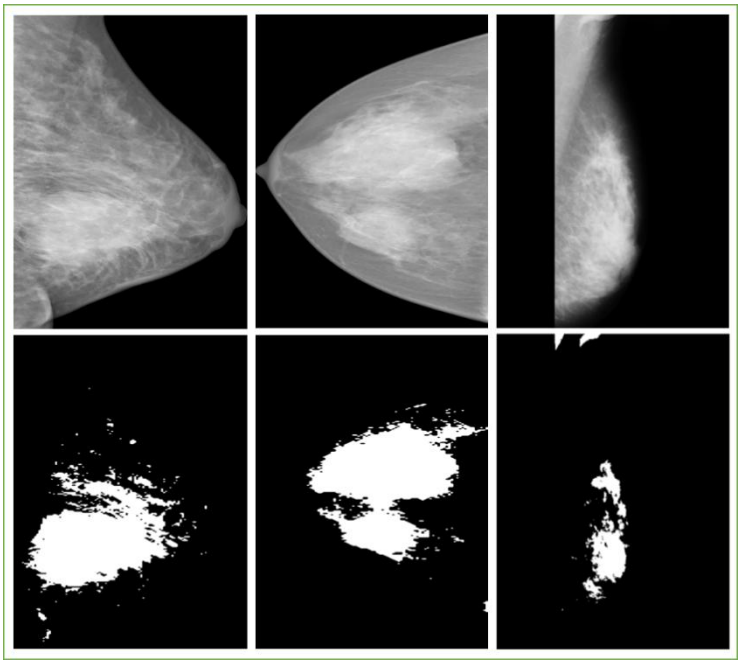

Figure 5.5: Sample mammograms with their segmentation masks

5.3.2 Implementation Details

This part explains the computing setup, software stack, and the main hyperparameters used during training and evaluation of SE-CRNN.

(a) Hardware:

- **GPU:** NVIDIA GeForce RTX 3080 (10–12 GB GDDR6)
    - Used for training both segmentation and classification branches.
- **CPU:** Intel Core i9 (8–16 logical cores, ~3.5–4.9 GHz)
    - Handles data loading, preprocessing, augmentation, and I/O.
- **System Memory:** 32 GB DDR4 RAM (around 3200 MHz)

- Used for buffering training and validation batches.

- **Storage:** NVMe SSD
    - Ensures fast dataset access and lower I/O latency.

This configuration is realistic for a high-end research workstation and can be matched by modern hospital servers if needed.

(b) Software and Frameworks:

- **Programming Environment:** Python 3.9
- **Deep Learning Framework:** TensorFlow 2.x with Keras API
- **Supporting Libraries:**
    - OpenCV and scikit-image — for image preprocessing and augmentation.
    - NumPy and pandas — for data handling and statistics.
    - Matplotlib — for plotting training curves and visual results.
- **Experiment Management:**
    - Custom training scripts with fixed random seeds to make the experiments as deterministic as possible.

(c) Training Hyperparameters: Unless otherwise mentioned, the following settings are used:

- **Input size:**
    - All images resized to 256 × 256, single-channel (grayscale), after bicubic interpolation and AHE-based normalization.
- **Batch size:**
    - 16 for segmentation (SUNet).
    - 32 for classification (TeResNet), depending on GPU memory.
- **Optimizer:**
    - Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-7}$.

- **Initial learning rate:**
    - 1×10^{-3}, reduced by a factor of 0.5 when the validation loss does not improve for 10 epochs (step-decay schedule).
- **Epochs and early stopping:**
    - Training runs for up to 100 epochs.
    - Early stopping is applied if the validation loss does not improve for 12 consecutive epochs, to avoid overfitting.
- **Joint loss balancing:**
    - For SE-CRNN joint training, the segmentation and classification losses are combined as: $\lambda_{\text{seg}} = 0.5$, $\lambda_{\text{cls}} = 0.5$.

Data augmentation (flips, slight rotations, contrast changes, and Gaussian noise) is applied on-the-fly during training to improve robustness see Table 5.6.

Table 5.6 — SE-CRNN Training Hyperparameters and Schedules

Hyperparameter	Segmentation Value	Classification Value	Notes
Input size	256 × 256 (grayscale)	256 × 256 (grayscale)	Bicubic resize + AHE
Batch size	16	32	Limited by GPU memory
Optimizer	Adam (LR = 1e-3)	Adam (LR = 1e-3)	LR halved on 10-epoch plateau
Epochs	up to 100	up to 100	Early stop on validation loss
Augmentation	flips, ±15° rotations, contrast jitter, Gaussian noise	same	Online, applied per mini-batch

5.3.3 Evaluation Metrics

To evaluate SE-CRNN in a clear and standard way, well-known metrics were used for both segmentation and classification [105]. The main formulas are referenced for completeness.

(a) Segmentation Metrics: For segmentation, the focus is on:

- **Intersection over Union (IoU) / Jaccard Index**
    - Measures how much overlap there is between the predicted mask P and the ground truth mask G.
    - Formally defined in

$$IoU = \frac{|P \cap G|}{|P \cup G|} \quad (5.25).$$

    - Higher IoU means the predicted and true tumour regions agree more closely.
- **Dice Similarity Coefficient (DSC)**
    - Another overlap measure, often more sensitive to small lesions than IoU as:

$$DSC = \frac{2|P \cap G|}{|P| + |G|} \quad (5.26).$$

    - It acts like a harmonic mean of precision and recall at the pixel level.

Both IoU and DSC are reported per class (benign, malignant) and also as mean values (mIoU, mDSC) over the classes.

(b) Classification Metrics

For classification, standard metrics are computed from true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN):

- **Accuracy** – overall proportion of correctly classified samples

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (5.27).$$

- **Precision** – how many of the predicted positive cases are actually positive

$$\text{Precision (Positive Predictive Value)} = \frac{\text{TP}}{\text{TP+FP}} \quad (5.28).$$

- **Recall (Sensitivity)** – how many of the truly positive cases are detected

$$\text{Recall (Sensitivity / True Positive Rate)} = \frac{\text{TP}}{\text{TP+FN}} \quad (5.29).$$

- **Specificity** – how well the model recognizes negative cases (benign or normal), given in $\text{Specificity (True Negative Rate)} = \frac{\text{TN}}{\text{TN+FP}}$ (5.30).

- **F1-score** – harmonic mean of Precision and Recall, balancing both

$$\text{F1} - \text{score} = \frac{2 \cdot PR \cdot SE}{PR + SE} \qquad (5.31)).$$

These metrics are reported for each dataset and, where applicable, averaged over cross-validation folds or repeated runs (Table 5.7).

Table 5.7: Metric definitions and interpretation

Metric	Formula	Interpretation
IoU	$\frac{\lvert P \cap G \rvert}{\lvert P \cup G \rvert}$	P \cap G
DSC	$\frac{2\lvert P \cap G \rvert}{\lvert P \rvert + \lvert G \rvert}$	P \cap G
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Overall correctness
Precision	$\frac{TP}{TP + FP}$	Positive prediction reliability
Recall	$\frac{TP}{TP + FN}$	Sensitivity to positives
Specificity	$\frac{TN}{TN + FP}$	Ability to detect negatives
F1	$\frac{2 \cdot PR \cdot SE}{PR + SE}$	Balance of precision & recall

Reporting protocol:

- For segmentation, reported the IoU and DSC per class and as mean values (mIoU, mDSC).
- For classification, reported Accuracy, Precision, Recall, Specificity, and F1-score.
- Where possible, computed 95% confidence intervals to show the variability and give a more reliable comparison between SE-CRNN and the baseline methods.

These metrics together give a broad and balanced view of how well the proposed framework performs in both localizing lesions and making correct diagnostic decisions.

5.3.4 Baseline Models and Comparison Setup

To judge how well the proposed SE-CRNN framework really performs, we compared it with several existed segmentation and classification baseline architectures. All baseline models were either re-implemented or taken from reliable reference code, and then trained on the same data splits, with the same preprocessing and augmentation pipeline. Optimization settings were also kept as similar as possible, so the comparison is fair.

(a) Segmentation Baselines: For the segmentation part, the following networks used as reference methods:

- **U-Net (standard)** – the classical encoder–decoder structure with skip connections, used as a strong baseline in medical image segmentation.
- **VU-Net (Vanilla U-Net variant)** – a slightly deeper U-Net-style model with minor architectural changes.
- **SegNet** – an encoder–decoder network that uses pooling indices to guide the upsampling path.
- **Connected-ResUNet (CRes-U-Net)** – combines residual blocks with the U-Net idea to improve gradient flow and feature reuse.
- **cGAN-CNN** – a conditional GAN-based segmentation approach, where a generator produces masks and a discriminator evaluates their realism.
- **RKO-UNet** – an optimized U-Net variant with architectural tweaks to reduce parameters while keeping performance.

(b) Classification Baselines: For classification, SE-CRNN (TeResNet branch) was compared against several standard and hybrid classifiers:

- **Plain CNN classifier** – a lightweight model with a few convolutional layers followed by fully connected layers, used as a simple baseline.

- **ResNet (ResNet-50 style)** – a deeper residual network backbone widely used for image classification tasks.
- **CNN + ResNet (feature fusion)** – features from a CNN and a ResNet are combined before the final decision layer.
- **CNN + SVM** – features extracted by a CNN are fed into a Support Vector Machine classifier instead of a fully connected softmax head.
- **cGAN-CNN hybrid classifier** – a classifier that uses features learned in a conditional GAN setting, adapted for benign vs malignant prediction.

All these models were trained under the same preprocessing, augmentation, and validation-based hyperparameter tuning strategy similar to SE-CRNN and details are presented in Table 5.8.

Table 5.8: Baseline Model Details with Input Size and approximate Parameters)

Model	Input Size	Approx. Params	Short Description
U-Net	256 × 256	7.8 M	Standard encoder–decoder
VU-Net	256 × 256	12.2 M	Deeper U-Net variant
SegNet	256 × 256	29.1 M	Encoder–decoder with pooling indices
CRes-U-Net	256 × 256	9.5 M	Residual blocks + U-Net structure
cGAN-CNN	256 × 256	14–20 M	Generator–discriminator segmentation pair
RKO-UNet	256 × 256	10.4 M	Optimized U-Net variant
ResNet (cls)	256 × 256	25.6 M	ResNet-50 style classification backbone
CNN (cls)	256 × 256	3.2 M	Lightweight CNN classifier
CNN + SVM	256 × 256	–	CNN feature extractor + SVM decision layer

(c) Reproducibility and Reporting: To keep the experiments reproducible, fixed random seeds were used wherever possible (data shuffling, augmentation, weight initialization). In addition, the seed values, hardware configuration, and software/library versions were recorded in an experiment log. For each reported quantitative result, the number of runs (n) used to compute the mean and standard deviation is clearly stated. Unless mentioned otherwise, major experiments are averaged over n = 3 independent runs with different seeds. In the results section,

numerical tables are complemented with visual outputs such as segmentation masks, confusion matrices, and ROC curves, so that the behaviours of SE-CRNN and the baselines can be understood both quantitatively and qualitatively.

5.4 Experimental Results

This section presents the experimental findings for the proposed SE-CRNN framework. We first describe how the model behaves during training, then report detailed segmentation and classification results. This is followed by comparisons with baseline methods, ablation studies, and a short discussion on efficiency and complexity.

5.4.1 Training Dynamics

During training, monitored both segmentation and classification branches together with their main metrics. For all three setups—CBIS-DDSM, INbreast, and the integrated dataset—SE-CRNN converged in roughly 80–100 epochs.

- **Segmentation branch (SUNet)**
    - The Dice Similarity Coefficient (DSC) and Intersection over Union (IoU) increased smoothly over epochs.
    - Fluctuations were small, which suggests stable learning.
    - The Dice curve usually reached a plateau a bit earlier than IoU, which is expected because Dice is more sensitive to boundary quality, especially for small lesions.
- **Classification branch (TeResNet)**
    - Classification accuracy grew steadily and crossed 94% within the first ~ 50 epochs for most datasets.
    - Validation curves stayed close to training curves, which indicates that overfitting was limited. The combination of SE blocks and BRNN seems to give a good regularizing effect.

5.4.2 Segmentation Results

(a) Quantitative performance: The segmentation evaluated using IoU and DSC, and compared SUNet with several baseline models (Table 5.9, 5.10 and 5.11) for benign, malignant, and overall (average) performance.

- For benign lesions, SUNet produced very strong overlaps, especially on CBIS-DDSM (IoU = 0.942, DSC = 0.969). Results on INbreast (IoU = 0.843, DSC = 0.914) and the integrated dataset (IoU = 0.826, DSC = 0.892) were also clearly ahead of the baselines.
- For malignant lesions, the advantages were even more obvious, with high IoU and DSC on all three datasets.
- When averaging across lesion types, SUNet achieved mIoU > 0.85 and mDSC > 0.91 consistently, which is higher than any of the competing methods.

Table 5.9: Benign class segmentation results (IoU / DSC)

Model	CBIS-DDSM (IoU / DSC)	INbreast (IoU / DSC)	Integrated (IoU / DSC)
SUNet	**0.942 / 0.969**	**0.843 / 0.914**	**0.826 / 0.892**
U-Net	0.735 / 0.867	0.808 / 0.923	0.765 / 0.846
VU-Net	0.895 / 0.946	0.724 / 0.790	0.659 / 0.736
SegNet	0.748 / 0.831	0.801 / 0.878	0.566 / 0.662
CRes-U-Net	0.781 / 0.875	0.712 / 0.810	0.650 / 0.782
cGAN-CNN	0.805 / 0.917	0.812 / 0.896	0.626 / 0.761
RKO-UNet	0.812 / 0.901	0.879 / 0.920	0.807 / 0.887

Table 5.10 — Malignant class segmentation results (IoU / DSC)

Model	CBIS-DDSM (IoU / DSC)	INbreast (IoU / DSC)	Integrated (IoU / DSC)
SUNet	**0.849 / 0.938**	**0.905 / 0.962**	**0.871 / 0.939**
U-Net	0.654 / 0.804	0.742 / 0.800	0.624 / 0.720
VU-Net	0.823 / 0.909	0.799 / 0.888	0.564 / 0.663
SegNet	0.685 / 0.805	0.721 / 0.827	0.668 / 0.785
CRes-U-Net	0.718 / 0.828	0.819 / 0.896	0.590 / 0.744
cGAN-CNN	0.842 / 0.915	0.880 / 0.928	0.817 / 0.893
RKO-UNet	0.863 / 0.930	0.845 / 0.890	0.805 / 0.862

Table 5.11 — Average segmentation performance (mIoU / mDSC)

Model	CBIS-DDSM (mIoU / mDSC)	INbreast (mIoU / mDSC)	Integrated (mIoU / mDSC)
SUNet	**0.895 / 0.954**	**0.878 / 0.940**	**0.853 / 0.917**
U-Net	0.694 / 0.836	0.775 / 0.861	0.694 / 0.785
VU-Net	0.861 / 0.931	0.762 / 0.839	0.615 / 0.699
SegNet	0.715 / 0.817	0.762 / 0.855	0.618 / 0.722
CRes-U-Net	0.751 / 0.854	0.770 / 0.855	0.624 / 0.764
cGAN-CNN	0.824 / 0.913	0.847 / 0.909	0.725 / 0.832
RKO-UNet	0.837 / 0.915	0.862 / 0.909	0.808 / 0.876

From these three tables (5.9, 5.10 and 5.11) we can see that SUNet consistently gives the best IoU and DSC, especially for malignant lesions, which are more critical from a clinical point of view. Figure 5.6 shows typical DSC growth and Figure 5.7 shows the IoU growth curves for segmentation of TeResNet. Together, they show that SE-CRNN learns in a stable and predictable way across all dataset configurations.

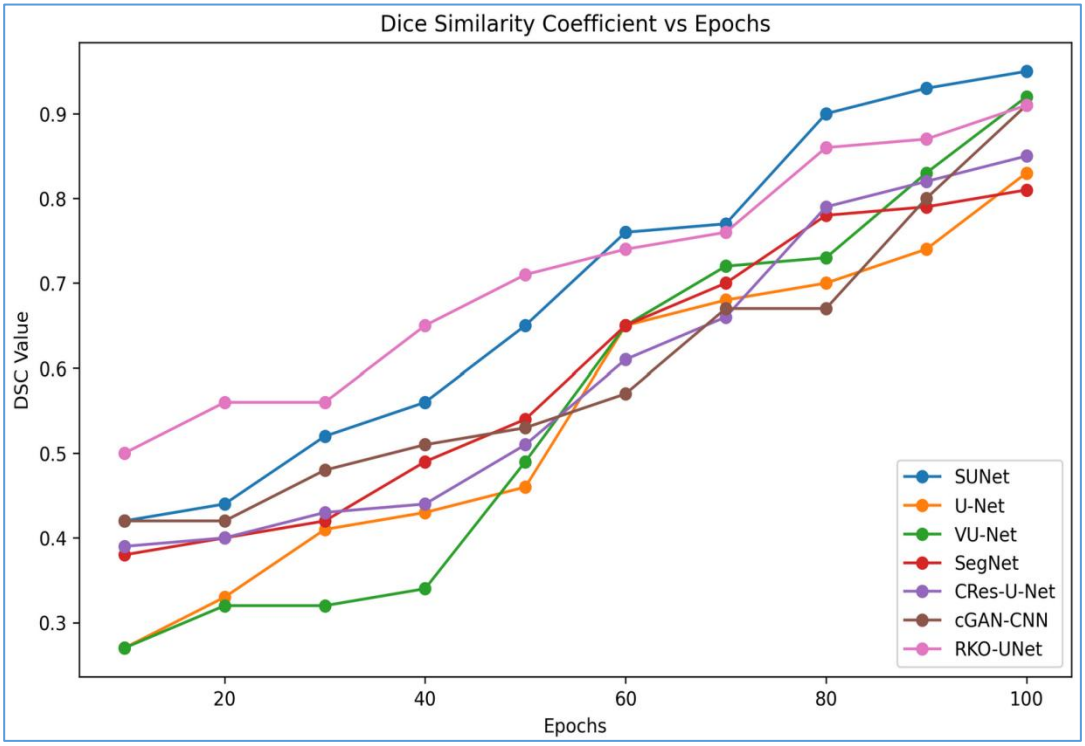

Figure 5.6: DSC growth of the segmentation models across training epochs

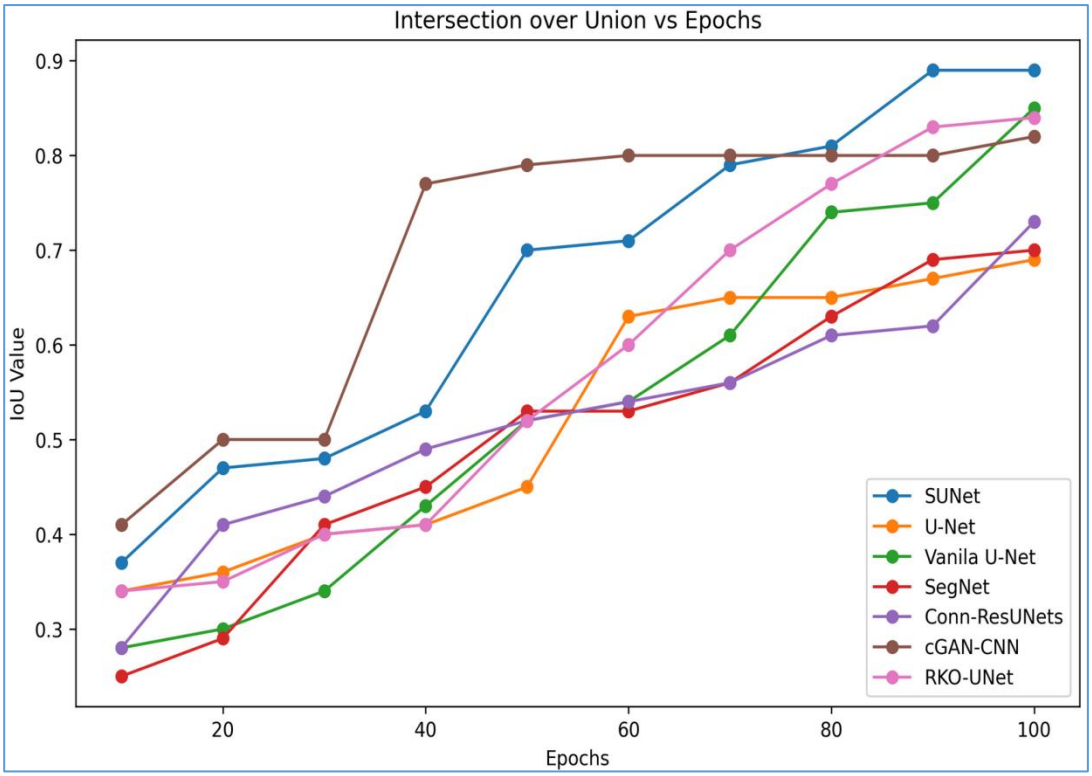

Figure 5.7: IoU growth of the segmentation models across training epochs

(b) Qualitative observations: Numbers alone cannot show whether the masks are visually convincing, so we checked the qualitative segmentation examples:

- On CC views (Figure 5.8), are the SUNet usually produced a sharp and smooth boundary that are matched the radiologist's masks well.
    - Benign lesions appeared as compact, rounded regions without unnecessary extra area.
    - Malignant lesions, often with spiculated or irregular shapes, were captured with high fidelity, and the predicted masks followed the complex contours closely.

On the MLO views (Figure 5.8), which are the more challenging due to the overlapping tissues SUNet maintained strong the alignment with the ground truth. Moreover, it detected the faint or partially obscured lesions more reliably than the traditional CNN-based models. Some of the failure cases also examined when:

- **Under-segmentation** mainly occurred when the lesion contrast was extremely low and the borders almost melted into surrounding tissue.

- **Over-segmentation** sometimes happened in very dense parenchyma, where non-tumorous tissue was mistakenly included in the mask.

Finally, these errors are not frequent, and they point to possible to improvements in preprocessing and attention tuning, rather than the fundamental problems in the architecture. The qualitative results from Figure 5.8 & 5.9 are further corroborate the quantitative findings, demonstrating that the SUNet is capable of producing clean, clinically meaningful segmentation masks.

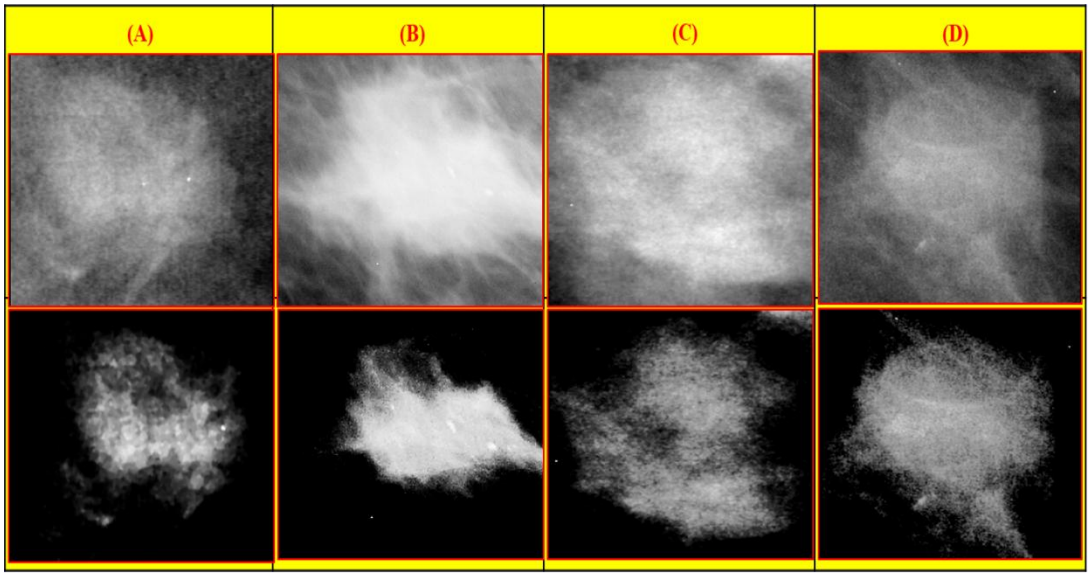

Figure 5.8: SUNet segmentation on CC-view images.

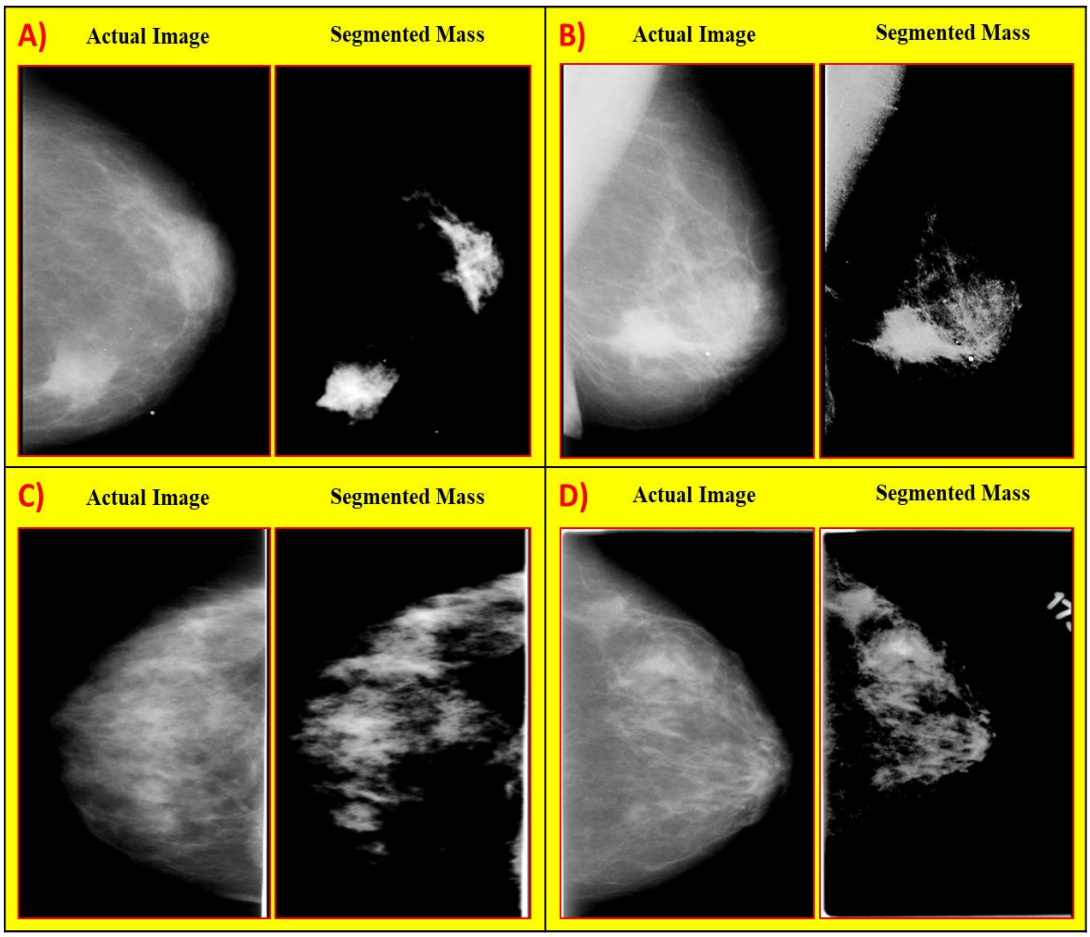

Figure 5.9: SUNet segmentation on MLO-view images.

5.4.3 Classification Results

For these classification tasks, we evaluated the TeResNet on the CBIS-DDSM, INbreast datasets, and the integrated datasets using sensitivity, specificity, accuracy, and F1-score. The primary objective was to assess the models are ability to detect the malignancies effectively while maintaining a low false-positive rates.

(a) Quantitative performance: On CBIS-DDSM dataset, TeResNet recorded the Sensitivity: 94.67%, Specificity: 96.65%, Accuracy: 95.68% and F1-Score: 95.67%. Similarly, on INbreast dataset, results stayed high as Sensitivity: 93.02%, Specificity: 95.24%, Accuracy: 94.38% and F1-Score: 94.12%. On the integrated dataset, TeResNet recorded the Sensitivity: 94.92%, specificity: 96.55%, Accuracy: 95.93% and F1-Score: 95.74%. Thus, even when the datasets with the differing imaging conditions are combined, to the model maintains a strong and well-balanced performance as presented with Table 5.12, 5.13 and 5.14.

Table 5.12 — CBIS-DDSM Classification Performance by TeResNet

Model	Sensitivity	Specificity	Accuracy	F1-Score
CNN+ResNet	93.83%	92.17%	93.15%	93.30%
CNN	87.34%	84.02%	85.32%	85.18%
CNN+SVM	84.12%	93.06%	88.71%	88.12%
ResNet	92.31%	90.02%	91.22%	91.27%
cGAN-CNN	89.74%	92.11%	91.25%	90.91%
RKO-UNet	88.37%	88.30%	92.51%	90.36%
TeResNet	**94.67%**	**96.65%**	**95.68%**	**95.67%**

Table 5.13 — INbreast Classification Performance by TeResNet

Model	Sensitivity	Specificity	Accuracy	F1-Score
CNN+ResNet	90.20%	88.46%	88.89%	89.32%
CNN	85.19%	92.02%	88.03%	88.46%
CNN+SVM	87.51%	90.57%	89.11%	89.72%
ResNet	83.67%	87.23%	85.86%	85.42%
cGAN-CNN	91.11%	95.35%	93.41%	93.18%
RKO-UNet	90.48%	92.68%	91.95%	91.57%
TeResNet	**93.02%**	**95.24%**	**94.38%**	**94.12%**

Table 5.14 — Integrated Dataset Classification Performance by TeResNet

Model	Sensitivity	Specificity	Accuracy	F1-Score
CNN+ResNet	87.51%	92.10%	90.59%	89.74%
CNN	83.33%	90.91%	86.73%	86.02%
CNN+SVM	84.01%	91.30%	87.88%	87.50%
ResNet	86.36%	89.41%	88.27%	87.86%
cGAN-CNN	92.65%	91.30%	92.36%	91.97%
RKO-UNet	89.55%	88.24%	90.21%	88.89%
TeResNet	**94.92%**	**96.55%**	**95.93%**	**95.74%**

The accuracy of the curves for each dataset results is shown in Figure 5.10 (CBIS-DDSM), Figure 5.11 (INbreast), and Figure 5.12 (integrated dataset). These accuracy graphs exhibit a smooth upward trend with the only minimal gaps between the training and validation accuracy. This pattern indicates the good generalization capability rather than the mere memorization.

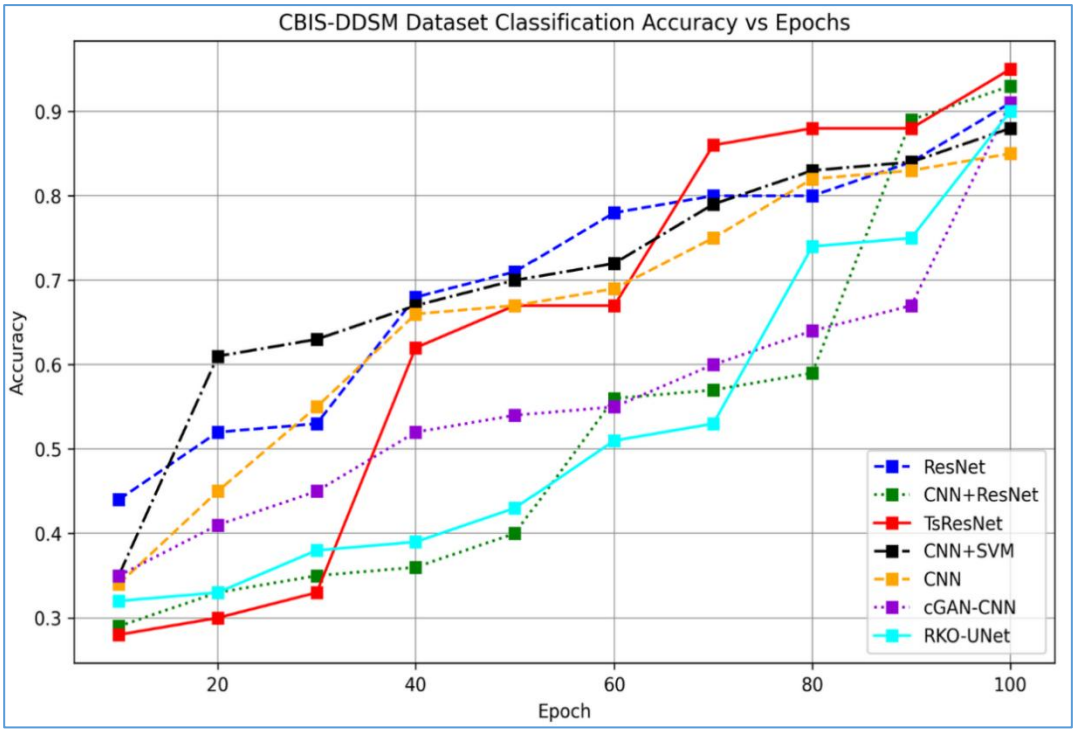

Figure 5.10: CBIS-DDSM Dataset Classification Accuracy growth vs Epochs

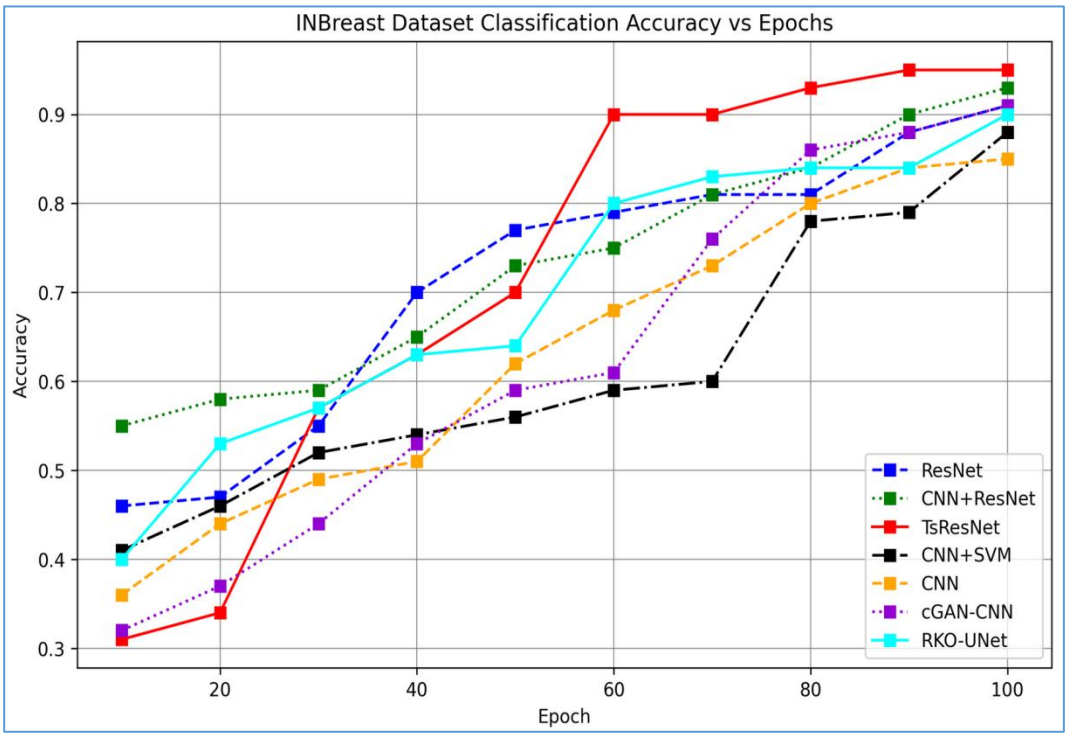

Figure 5.11: INBreast Dataset Classification Accuracy growth vs Epochs

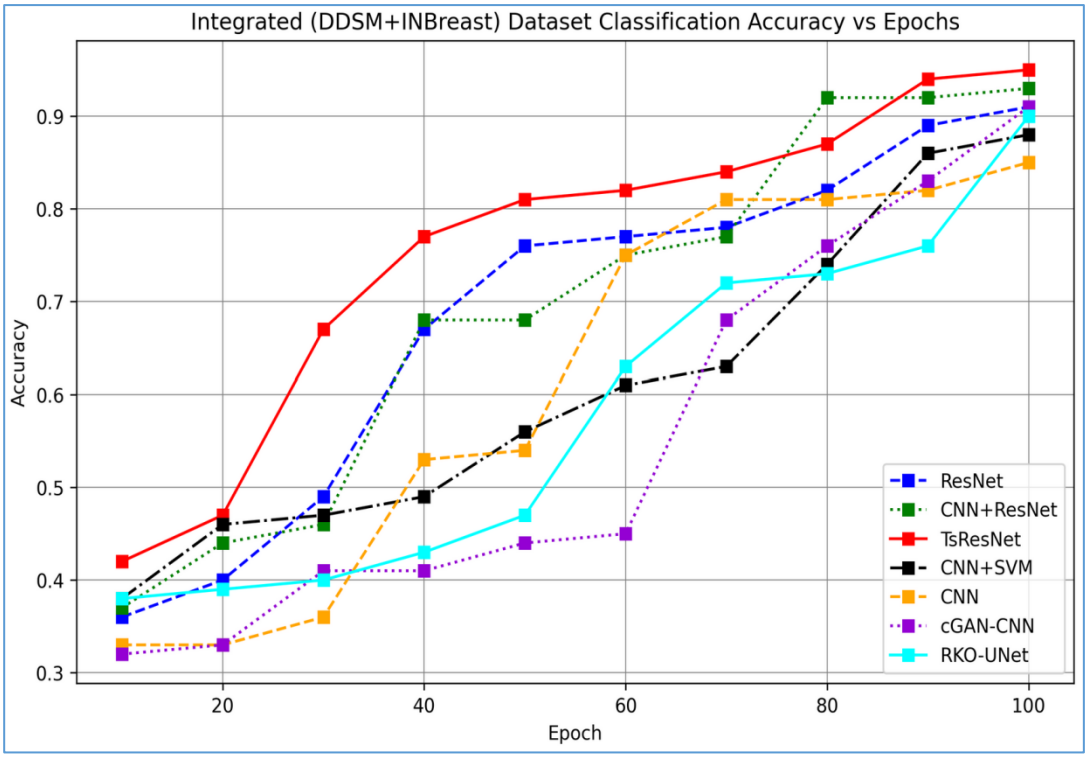

Figure 5.12: Integrated (DDSM+INBreast) datasets classification accuracy growth over Training epochs

5.4.4 Comparative Analysis vs. Baselines

The purpose of this comparative analysis is to determine whether the dual-branch design of SE-CRNN (SUNet + TeResNet) offers measurable advantages over traditional or hybrid architectures.

(a) Segmentation Comparison: Across all lesion types and datasets:

- SUNet outperformed U-Net, VU-Net, SegNet, CRes-U-Net, cGAN-CNN, and RKO-UNet.
- The performance gains were most pronounced for malignant lesions, whose boundaries tend to be small, irregular, and heavily affected by dense tissue.
- The SE blocks appear to contribute significantly by recalibrating feature channels and maintaining focus on tumour-related regions, resulting in improved boundary preservation.

Figure 5.6 and 5.7 (graphs of DSC and IoU) shows SUNet consistently achieving the highest performance across domains, reinforcing its reliability and stability for clinical segmentation tasks.

(b) Classification comparison:

- TeResNet achieved higher sensitivity and specificity than plain CNNs, ResNet alone, and hybrids like CNN+SVM or cGAN-based classifiers.
- The integration of a residual backbone with BRNN-based temporal modelling enables the network to capture inter-view relationships and subtle temporal or feature patterns that conventional single-image models are unable to detect.

In classification plots (Figure 5.10, 5.11 and 5.12) of different datasets, TeResNet exhibits a large, nearly circular profile across sensitivity, specificity, accuracy, and F1-score, indicating a well-balanced overall performance. Other models show more irregular shapes, which indicates trade-offs (i.e., high sensitivity but lower specificity).

(c) Interpretation

- **SUNet**: main advantage is strong spatial boundary modelling with SE-driven channel attention.
- **TeResNet**: main advantage is use of temporal/context information on top of ResNet features.

Together, they make SE-CRNN a coherent framework that improves not only raw metrics but also cross-dataset robustness and interpretability, both of which are crucial for real clinical use.

5.4.5 Ablation Studies

To understand the contribution of each major component, ablation experiments conducted and results are presented in Table 5.15, with three reduced configurations:

1. **Without SE block** – SE removed, so no channel-wise spatial recalibration.
2. **Without BRNN** – temporal modelling removed; only spatial CNN features used.
3. **Without SE and BRNN** – only the basic CNN backbone with standard heads.

(a) Observations

- **Without SE block**
    - IoU and DSC dropped by about 3–4% on average.
    - Visual masks showed blurrier boundaries and more false positives in dense tissue.
    - This confirms that SE is important to highlight the most discriminative channels.
- **Without BRNN**
    - Classification performance decreased; sensitivity went down by roughly 5%.
    - The model became weaker at recognizing subtle malignant patterns which need contextual or multi-view information.
- **Without both SE and BRNN**
    - Performance in both segmentation and classification dropped close to a standard CNN/ResNet setup.
    - This shows that the gains of SE-CRNN really come from the combination of spatial recalibration and temporal modelling see Table 5.15.

Table 5.15 — Ablation performance deltas (relative to full SE-CRNN)

Configuration	CBIS-DDSM (IoU / DSC)	INbreast (IoU / DSC)	Integrated (IoU / DSC)	Accuracy (Cls.)	Sensitivity (Cls.)	F1-Score (Cls.)
Full SE-CRNN	0.895 / 0.954	0.878 / 0.940	0.853 / 0.917	95.90%	94.90%	95.70%
w/o SE Block	0.864 / 0.922	0.842 / 0.902	0.819 / 0.885	94.20%	91.50%	92.90%
w/o BRNN	0.879 / 0.935	0.861 / 0.918	0.828 / 0.893	92.80%	89.80%	91.00%
w/o SE + BRNN	0.812 / 0.891	0.803 / 0.876	0.774 / 0.852	90.10%	86.70%	88.30%

Without SE, contours become less sharp and without BRNN, some lesions are misclassified and poorly segmented.

5.4.6 Efficiency and Complexity

Besides accuracy, a model intended for clinical deployment must also be computationally feasible. Therefore, we compared the SE-CRNN with baseline models (Table 5.16) in terms of parameters, FLOPs, and training time per epoch on an NVIDIA RTX 3080.

(a) Observations

- **Model size**
    - SE-CRNN has more parameters than a plain U-Net or ResNet due to SE and BRNN, but remains smaller than heavier models like SegNet or some GAN-based methods.
- **FLOPs**
    - Compared to U-Net, FLOPs increase by about 15%, mainly because of SE blocks and recurrent connections.
    - Still, the computational load is clearly below that of cGAN-based architectures.
- **Training time**
    - SE-CRNN: about 62 seconds per epoch on RTX 3080.
    - U-Net: ~48 s/epoch.
    - cGAN-CNN: ~95 s/epoch.

So, SE-CRNN sits between light and heavy models, but offers a much better accuracy–complexity trade-off.

Table 5.16: Computational profile of SE-CRNN vs. Baselines

Model	Params (M)	FLOPs ($\times 10^9$)	Time per Epoch (s)	Accuracy (Cls.)	mDSC (Seg.)
U-Net	7.8	22.5	48	91.20%	0.836
SegNet	29.1	38.6	88	89.70%	0.817
CRes-U-Net	9.5	24.3	55	92.10%	0.854

cGAN-CNN	18.2	42.7	95	92.30%	0.913
RKO-UNet	10.4	27.1	67	93.50%	0.876
SE-CRNN	**12.7**	**25.9**	**62**	**95.90%**	**0.917**

5.5 Discussion

The extensive experiments on CBIS-DDSM, INbreast, and the integrated dataset allowed me not only to compare numbers for SE-CRNN but also to understand why the framework works well, how far it can generalize, and what this might mean in a real clinical setting. In this section, we bring together the main findings, relate them to the architectural choices, and then discuss the strengths, limitations, and possible clinical impact of SE-CRNN.

5.5.1 Key Findings

All of these experimental results show that SE-CRNN clearly performs better than the competing segmentation and classification methods.

- **Segmentation (SUNet branch):** Across all datasets, SUNet achieved a mean Dice Similarity Coefficient around 0.919 and a mean Intersection over Union around 0.854. These values are higher than those of classical U-Net, VU-Net, and cGAN-based approaches. This suggests that adding spatial attention through SE blocks helps the model preserve boundaries more accurately, especially for malignant lesions with irregular and complex edges.
- **Classification (TeResNet branch):** For classification, TeResNet reached an accuracy of about 95.9% and a sensitivity of around 94.9%, while keeping a good balance between detecting malignant cases and avoiding too many false alarms in benign cases [106]. The F1-scores above 95% on all datasets indicate that the classifier is both reliable and robust, even when the data distribution changes.

5.5.2 Why SE-CRNN Works

The good performance of SE-CRNN does not come from a single component, but from the way several modules are combined in a synergistic way.

SE block – refined spatial features: The SE block recalibrates channel-wise feature responses. In simple words, it increases the influence of channels that are strongly

related to tumour regions and reduces the effect of channels dominated by normal tissue or noise. This highlights faint lesion boundaries, improves detection of small or subtle tumours, and reduces confusion caused by dense background tissue. As a result, the segmentation masks from SUNet tend to follow the true lesion contours more closely than those from plain CNN-based backbones.

BRNN – temporal and sequential context: The Bidirectional RNN (BRNN) module models forward and backward dependencies in feature sequences. This allows the network to link information across multiple views (e.g., CC and MLO), and to exploit relationships across hierarchical feature maps. By looking at the sequence from both directions, the network becomes more sensitive to subtle malignant patterns that may not be obvious in a single static view. This contributes mainly to the higher sensitivity and robustness of TeResNet.

Joint architecture – shared backbone: Both SUNet (segmentation) and TeResNet (classification) are built on a shared spatial–temporal backbone (CNN + SE + BRNN). This joint design means that:

- features are useful for segmentation also benefit classification, and
- The classification supervision helps the backbone learn features that also improve segmentation.

This mutual reinforcement leads to better performance than single-task models or purely spatial pipelines.

5.5.3 Generalization and Robustness

One of the most important outcomes of this work is that SE-CRNN shows consistent behaviours across heterogeneous datasets.

- On CBIS-DDSM (film-derived mammograms), the framework maintained high segmentation and classification metrics despite variations in film digitization, image quality, and acquisition protocol.
- On INbreast (digital FFDM), performance remained strong even though the dataset is smaller and has different imaging characteristics. This suggests that SE-CRNN can adapt to modern digital systems without needing heavy redesign.

- On the integrated dataset (CBIS-DDSM + INbreast), SE-CRNN did not suffer from the strong performance drop that many models show under domain shift. Instead, it delivered stable results across both domains.

These observations indicate that the framework is robust to variations in scanner type, resolution, and patient population. This is encouraging for real-world screening environments, where data are rarely clean, uniform, or collected from a single source.

5.5.4 Clinical Implications

For real deployment in hospitals, a model should not only be accurate, but also reasonably interpretable and easy to integrate into the clinical workflow. SE-CRNN contributes to these aspects in several ways. With dual combination of precise segmentation and reliable classification, SE-CRNN can be used as part of a computer-aided diagnosis (CAD) system, for example as a second reader:

- The segmentation masks assist in treatment planning (e.g., defining tumour extent for surgery or radiotherapy).
- The classification outputs help in risk stratification and may support decisions such as biopsy recommendation or follow-up interval.

In dealing with screening issues, this tool can potentially reduce reading time, help priorities suspicious cases, and maintain high sensitivity for early-stage cancers.

5.5.5 Limitations

Although SE-CRNN shows promising results, there are still several limitations (Table 5.17) that should be recognized and may guide future work.

Computational cost: The use of SE blocks and BRNNs increases model complexity compared to lightweight CNNs. Training and inference demand more GPU memory and time. In high-resource centres this is manageable, but in low-resource clinical settings deployment might be challenging without additional model compression or hardware support.

Dataset and modality scope: All experiments in this chapter are based on mammography only. While SE-CRNN generalizes well across CBIS-DDSM and INbreast, it has not yet been tested on other imaging modalities such as breast MRI,

ultrasound, or tomosynthesis. Its real performance on those modalities is therefore still unknown.

Lack of multi-modality integration: In real clinical practice, radiologists often combine information from several modalities (e.g., mammogram + ultrasound + MRI). The current version of SE-CRNN processes only mammograms. Without multi-modality fusion, some complex diagnostic scenarios may still require additional manual interpretation.

Remaining interpretability challenges: Even with attention maps, SE-CRNN is still, in many ways, a black-box model. The deeper internal reasoning and feature interactions are not fully transparent. See Table 5.17 More advanced Explainable AI (XAI) tools could be integrated in the future to provide richer explanations and further increase clinician confidence.

Table 5.17: Summary of Strengths and Limitations of SE-CRNN

Aspect	Strengths	Limitations
Segmentation (SUNet)	High mDSC (~0.919); good boundary preservation	Heavier than plain U-Net; higher computation needed
Classification (TeResNet)	Accuracy ≈ 95.9%; balanced sensitivity/specificity	Currently evaluated only on mammograms
Generalization	Stable across CBIS-DDSM, INbreast, and integrated data	Not yet tested on other imaging modalities
Interpretability	Attention maps provide visual cues for decisions	Deeper decision logic remains largely opaque

5.6 Future Work

Although SE-CRNN already shows clear gains in both segmentation and classification, there are still several directions where the work can be improved and extended. In this section we outline the main lines of future research (Table 5.18) that can consider important in terms of methodology and clinical use.

1) Multimodal Integration with MRI and Ultrasound: In routine breast cancer diagnosis, radiologists do not rely only on mammograms. MRI is often used for dense breasts and high-risk patients, and ultrasound is commonly used to differentiate cystic and solid masses and to guide procedures. A natural next step for

SE-CRNN is therefore to move from single-modality to multimodal learning. In this extension, SE-CRNN would take mammogram, MRI, and ultrasound inputs and learn joint spatio–temporal features across these modalities. By fusing heterogeneous but complementary information, the system could obtain a more complete description of lesion shape, internal texture, and temporal behaviours.

2) Semi-Supervised Learning with Limited Labels: High-quality pixel-level and case-level annotations are expensive and time-consuming, because they require experienced radiologists. In many hospitals, especially in low-resource settings, only a small portion of mammograms can be labelled in detail. To reduce this dependency on fully labelled data, SE-CRNN can be extended with semi-supervised or weakly supervised strategies, for example:

- **Pseudo-labelling**: use confident predictions on unlabelled images as temporary labels and refine them over time.
- **Consistency regularization**: enforce that the model gives similar outputs under different perturbations of the same image.
- **Teacher–student frameworks**: train a large "teacher" model first, and then transfer knowledge to a smaller "student" using both labelled and unlabelled data.

Such techniques would allow SE-CRNN to make better use of large pools of unlabelled mammograms, aiming for almost the same performance as a fully supervised model, but with fewer expert annotations.

3) Model Compression for Practical Deployment: Even though SE-CRNN shows a good balance between performance and complexity, it is still heavier than simple CNN baselines. In real-world deployment, especially on portable devices or on hospital systems without powerful GPUs, this can be a problem. Future work will therefore explore model compression techniques, such as:

- **Pruning**: remove redundant filters and weights that contribute little to the final prediction.
- **Quantization**: store and compute with reduced-precision parameters (for example, 8-bit or mixed precision instead of full 32-bit floats).

- **Knowledge distillation**: train a smaller "student" network to mimic the outputs or feature distributions of the full SE-CRNN.

The goal is to reduce model size and inference time while keeping accuracy as high as possible, so that SE-CRNN can run close to real time even on modest hardware.

4) Extension to Tumour Staging and Survival Prediction

So far, SE-CRNN focuses on lesion segmentation and binary malignancy classification. In clinical oncology, however, doctors are also interested in Tumour staging (e.g., TNM categories), and Survival or recurrence prediction and overall prognosis. A promising extension is to integrate clinical metadata (age, hormonal receptor status, histology, treatment details) and possibly genomic or molecular markers with imaging features learned by SE-CRNN. This would allow the framework to perform: Staging (e.g., mapping image and clinical features to TNM stages) and Risk and survival prediction (e.g., predicting recurrence risk or survival time ranges). In this way, SE-CRNN can move from being only a diagnostic tool to a prognostic and decision-support tool, which is more in line with modern precision medicine.

Table 5.18: Future Roadmap for SE-CRNN

Goal	Planned Methodology	Expected Impact
Multimodal integration (MRI / US)	Joint fusion of mammogram, MRI, and ultrasound streams	Richer and more reliable tumour characterization
Semi-supervised learning	Pseudo-labelling, consistency loss, teacher–student models	Lower need for fully annotated datasets
Model compression	Pruning, quantization, knowledge distillation	Faster, lighter models suitable for low-cost systems
Staging & survival prediction	Fusion of imaging, clinical, and genomic information	Added prognostic value and support for treatment planning

These directions together define a medium- to long-term plan to make SE-CRNN more scalable, efficient, and clinically useful in different environments.

5.7 Chapter Summary

In this chapter, we presented SE-CRNN, a spatiotemporal deep learning framework designed for joint breast lesion segmentation and classification in mammography. The chapter started by explaining the motivation to combine spatial precision with temporal modelling, since traditional CNNs focus mainly on static spatial patterns. SE-CRNN addresses this by introducing two main components:

- **SUNet** – a U-Net style segmentation network with Squeeze-and-Excitation (SE) blocks, which improves channel-wise attention on tumour-related features and leads to high segmentation accuracy (around mDSC ≈ 0.919, mIoU ≈ 0.854).
- **TeResNet** – a classification head that merges residual learning with Bidirectional Recurrent Neural Network (BRNN)-based temporal modelling, achieving high accuracy (about 95.9%) and sensitivity (around 94.9%) across multiple datasets.

Extensive experiments on CBIS-DDSM, INbreast datasets, and an integrated dataset showed that the SE-CRNN:

- Outperforms strong baselines such as U-Net, SegNet, ResNet, and cGAN-based methods in both segmentation and classification.
- Remains robust under heterogeneous imaging conditions, including different resolutions, acquisition protocols, and breast densities.

Ablation studies further the clarified into:

- The SE blocks mainly improve segmentation, especially fine lesion boundaries and small structures.
- The BRNN module primarily enhances the classification performance by capturing the temporal or multi-view contextual information. When combined, SE and BRNN yield synergistic improvements, surpassing the benefits are provided by either module individually.

The efficiency analysis shows that the SE-CRNN is moderately more demanding than the ultra-lightweight models; however, its accuracy–complexity trade-off

remains favourable, particularly when compared to with more computationally intensive the GAN-based architectures. From a clinical standpoint, are proposed framework offers the two key advantages:

- Accurate lesion delineation, supporting the more informed treatment planning, and
- Reliable malignancy prediction, aided by the SE- and BRNN-derived from attention maps that provide a useful degree of visual interpretability for radiologists.

Several limitations were also acknowledged, including the exclusive focus on the mammography, the computational overhead of the model, and the absence of the multimodal data integration. These constraints are naturally motivated by the outlined future works, such as multimodal fusion, semi-supervised learning approaches, models are compression strategies, and extensions toward the prognostic modelling. In this summary, this chapter demonstrates that the SE-CRNN constitutes a strong foundation for the spatiotemporal deep learning in breast imaging and establishes a clear trajectory toward more advanced and clinically relevant CAD systems are to be explored in the subsequent chapters and future studies.

CHAPTER – 6 HMST-LITE FRAMEWORK FOR EARLY BREAST CANCER STAGE CLASSIFICATION

6.1 Introduction

6.1.1 Clinical Background and Motivation

Breast cancer is still one of the main causes of cancer-related death in women across the world [2 and 4]. A large number of patients are unfortunately diagnosed only in advanced stages, when treatment is less effective and the chance of long-term survival is lower. Clinical studies shown in figure 6.1, women diagnosed at stage 0 have a five-year survival rate above 90%, while this value often falls below 40% for stage III or IV cases [107]. These numbers highlight that how important early detection is for improving survival, reducing recurrence, and planning more suitable treatments.

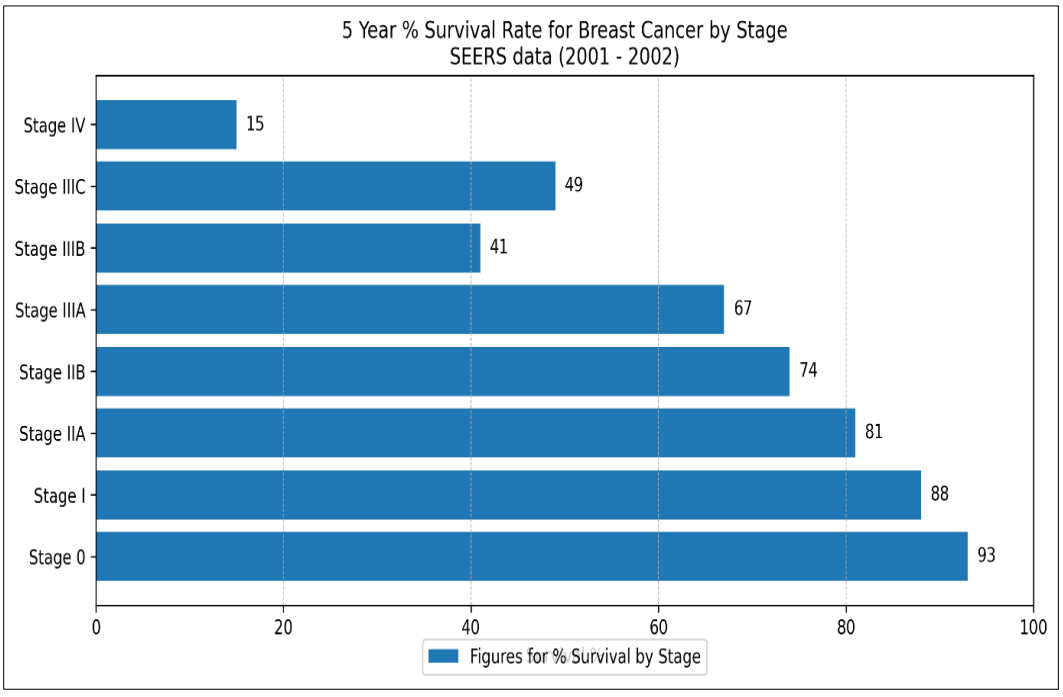

Figure 6.1: Breast cancer survival rate trend by stage (Stage I–IV).

In current practice, diagnosis and staging mainly depend on radiology and histopathology, which is usually reported under standard systems such as the AJCC TNM staging framework [108]. Radiological imaging especially mammography and MRI is very helpful for screening and for finding suspicious regions. However, these images do not always show the fine-level structural and micro-environment details

of the tumour. On the other hand, histopathological staging on whole-slide images (WSIs) is still considered the gold standard [109], but it is labour-intensive, time-consuming, and affected by inter-observer variability. In difficult or borderline cases, pathologists do not always agree on the grade or stage, and reported agreement levels. The variability, together with the manual workload, creates a strong need for automatic, standardized, and robust computational methods to support breast cancer detection and staging [19].

6.1.2 Computational Challenges

Even though deep learning has brought significant progress in medical image analysis, automatic staging of breast cancer from histopathology WSIs is still a very challenging task. The main difficulties are summarized below as follows.

1) Extremely high-resolution WSIs and computational limits: Whole-slide images can easily exceed 100,000 × 100,000 pixels, so direct end-to-end processing is practically impossible with normal GPU memory. Therefore, most methods adopt a patch-based approach: they sample small tiles from the slide and process them individually. However, this creates a new problem—how to rebuild the global context and tissue structure from many small, local patches. If this global view is not properly recovered, the staging decision may miss important patterns.

2) Multi-scale dependencies: Cancer patterns appear at different spatial scales at the same time:

- **Cell level** – nuclear size, shape irregularity, mitotic figures.
- **Tissue level** – stromal invasion, duct structures, overall tumour boundary.

A good staging model should combine both levels. Traditional single-scale or fixed-resolution models often ignore these cross-scale relations. As a result, they may learn incomplete or context-poor features and fail to capture all staging cues.

3) Dataset variability and domain shift: WSI datasets from different hospitals and scanners can vary a lot due to: the different staining protocols, scanner hardware and image resolution, and annotation style and quality [110]. Models trained on one cohort often show a clear performance drop when applied directly to another cohort. This domain heterogeneity is a serious obstacle for real deployment in multi-centre clinical environments as shown in Figure 6.2.

4) Interpretability gap in AI models: Many deep learning systems behave like black boxes: they provide a stage prediction but do not clearly show *why* they reached that decision [54]. In medicine, however, clinicians need transparent and interpretable reasoning before they can trust and regularly use an AI system. Lack of interpretability slows down acceptance and limits the practical use of these models, as summarized in Table 6.1.

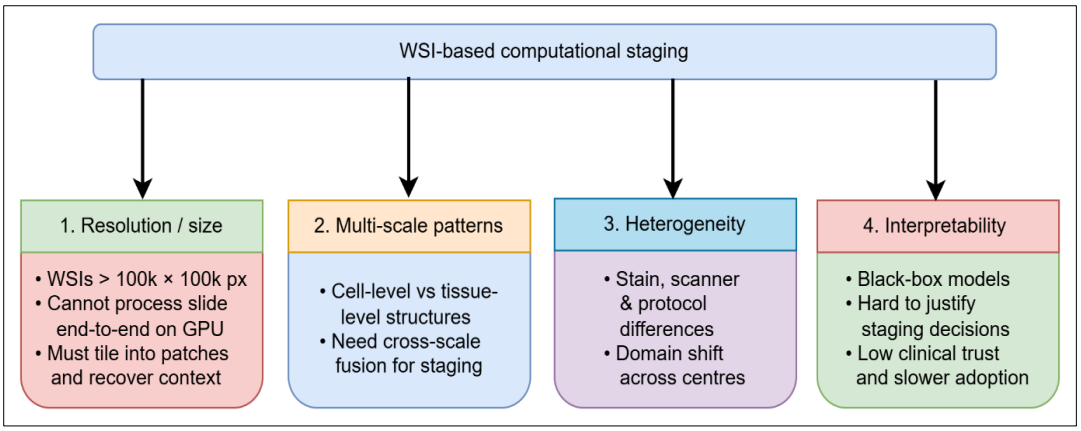

Figure 6.2: Illustration of key challenges in WSI-based computational staging (resolution, scale, heterogeneity, interpretability).

Table 6.1: Main Computational Challenges in Breast Cancer WSI Analysis

Challenge	**Example Issue**	**Clinical Impact**
High-resolution WSIs	> 100k × 100k pixels per slide	Very high memory and computation requirements
Multi-scale dependencies	Cell-level vs tissue-level patterns	Important staging cues may be missed at one scale
Dataset variability	Staining, contrast, annotation style	Poor generalization across hospitals/cohorts
Interpretability gap	Black-box predictions	Lower clinical trust and slower adoption

6.1.3 Objectives and Scope

This chapter tackles the above issues by proposing HMST-Lite, a lightweight transformer-based framework for early breast cancer stage classification from WSIs [111]. The main idea is to design a model that is accurate, interpretable, and efficient enough for possible deployment in real clinical settings. The main objectives are:

1. **Design a lightweight dual-scale Vision Transformer (ViT):** HMST-Lite uses two separate transformer branches, working at 10× and 20× magnifications. This design imitates how human pathologist's work: they zoom out to look at the global tissue organization and zoom in to inspect cellular details [27].

2. **Improve accuracy and interpretability together:** The model fuses the class tokens from both magnification branches to capture complementary global and local features. At the same time, it produces attention heatmaps that highlight important regions. These maps can be compared with known histological markers which make the predictions more understandable and easier to verify by experts.

3. **Support efficient deployment:** HMST-Lite is intentionally kept lightweight, with fewer parameters and reduced inference time compared to standard ViT architectures. The goal of this framework is presented in Figure 6.3, specifies to enable the use in resource-limited clinical environments, such as smaller hospitals or shared GPU servers [28].

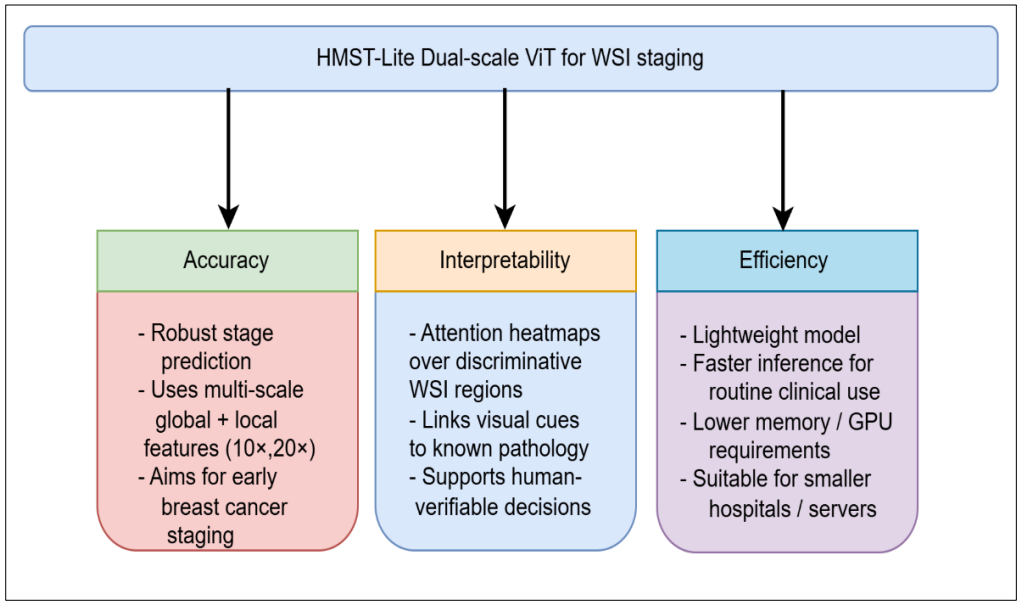

Figure 6.3: Conceptual overview of HMST-Lite goals – accuracy, interpretability, and efficiency.

6.2 HMST-Lite Framework

6.2.1 Architecture Overview

In this work, HMST-Lite (Figure 6.4) is designed as a dual-branch vision transformer that analyses breast histopathology images at two magnification levels:

- **10× (low magnification)** – to capture global, tissue-level structure.
- **20× (high magnification)** – to capture local, cellular-level details [60].

This two-scale design follows the practical behaviours of pathologists, who routinely move between low and high magnification to combine overall context with fine morphology [17]. Compared with conventional CNN-based staging models and very heavy ViT architectures, HMST-Lite focuses on efficiency and interpretability. It is specifically tailored for resource-limited clinical environments such as regional hospitals and diagnostic centres where GPU memory and computation are restricted [15 and 32]. The model reduces parameter count and inference time, but still aims to keep staging performance at a clinically useful level.

Each magnification branch acts as an independent ViT encoder, and the overall workflow has three main steps:

1. **Patch extraction and tokenization**
    - The WSI tile is split into non-overlapping patches at a given magnification.
    - Each patch is converted into a vector embedding.
2. **Transformer encoding**
    - Multi-head self-attention and feed-forward blocks learn within-scale (intra-scale) dependencies among patches.
3. **Aggregation via class token**
    - A special class token collects a compact representation of all patches within each branch.

After encoding, the class-token representations from the 10× and 20× branches are concatenated and passed to a shallow MLP for stage classification [112]. This

lightweight fusion avoids expensive cross-attention between branches and keeps the architecture simple, scalable to new datasets, and easier to interpret.

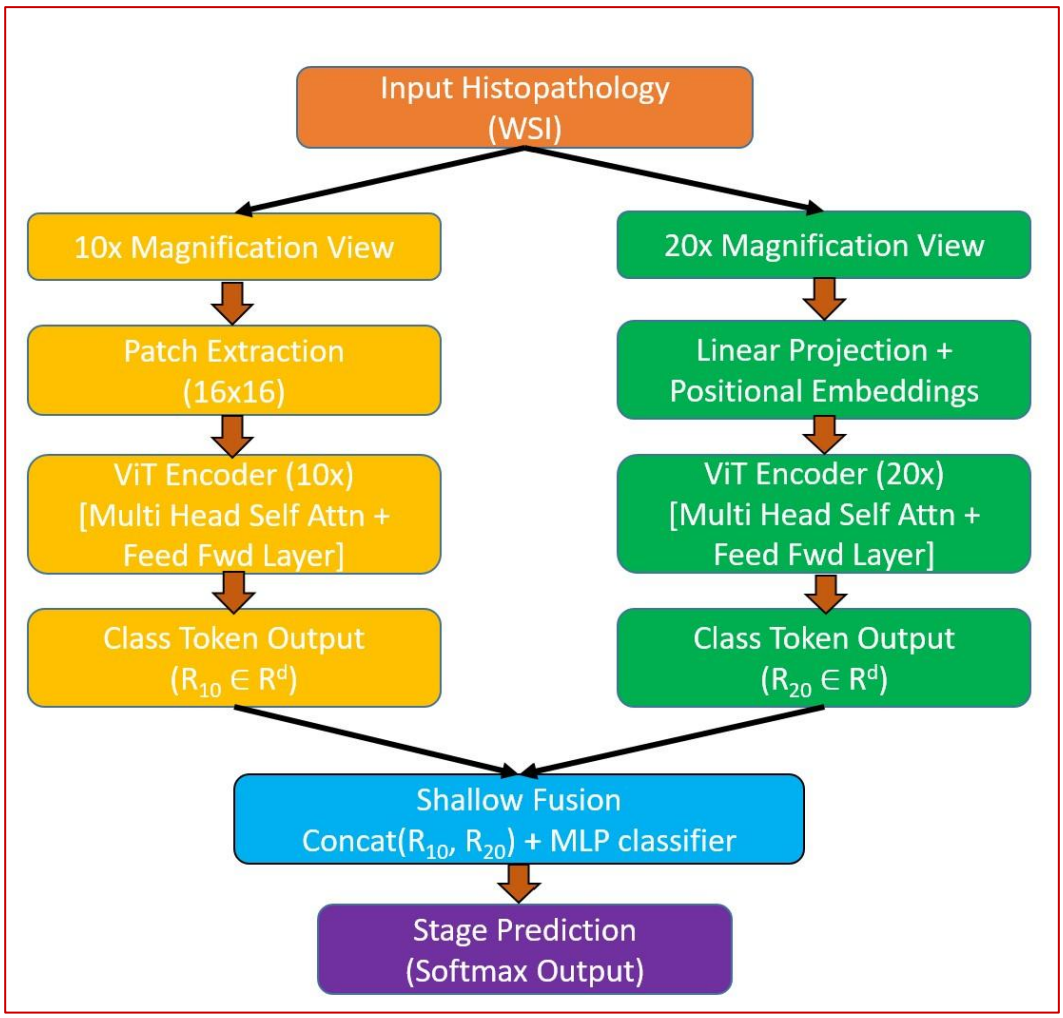

Figure 6.4: Block diagram of the HMST-Lite architecture.

6.2.2 Patch Extraction and Tokenization

Whole-slide images are extremely large (often above 100,000 × 100,000 pixels), so direct processing is not realistic [88]. HMST-Lite therefore works on WSI tiles, which are further divided into non-overlapping square patches of fixed size, depending on the magnification level as shown in Figure 6.5 and 6.6. If an input tile at scale $s$shows dimensions(H_s, W_s), the total number of patches N_s is computed as

$$N = \frac{H \times W}{P^2} \quad (6.1)$$

Each patch $\mathbf{x}_i \in \mathbb{R}^{P \times P \times C}$, with patch size P and C channels (typically $C = 3$for RGB), is then flattened into a 1-D vector (Ritesh Maurya et al., 2024):

$$p_i \in \mathbb{R}^{P^2C} \quad (6.2)$$

A linear projection maps this flattened vector into a D-dimensional embedding space:

$$e_i = W_E p_i + b_E \quad (6.3)$$

where W_E is the learnable projection matrix and b_E is a bias term. To perform classification, a trainable class token z_{cls} is prepended to the patch embeddings. It acts as a global summary of all patches within that branch. The resulting input sequence is

$$Z_0 = [z_{cls};\ e_1;\ e_2;\ \ldots;\ e_{N_s}] \quad (6.4)$$

To preserve spatial information, positional encodings P are added element-wise:

$$\tilde{Z}_0 = Z_0 + P \quad (6.5)$$

The final token sequence $\tilde{Z}_0$ is then passed into the transformer encoder, where multi-head self-attention learns patch-to-patch relationships at the given scale.

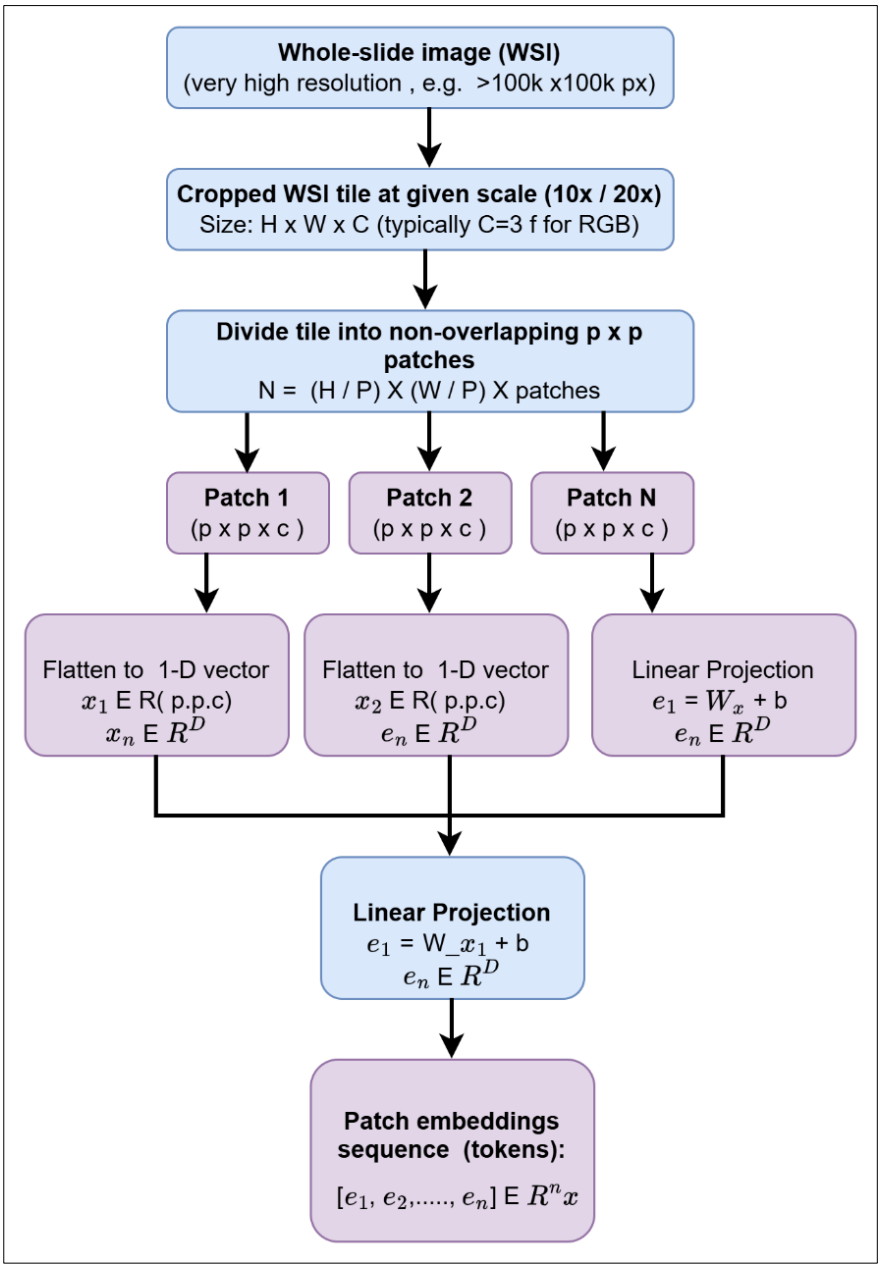

Figure 6.5: Patch extraction and tokenization pipeline in HMST-Lite.

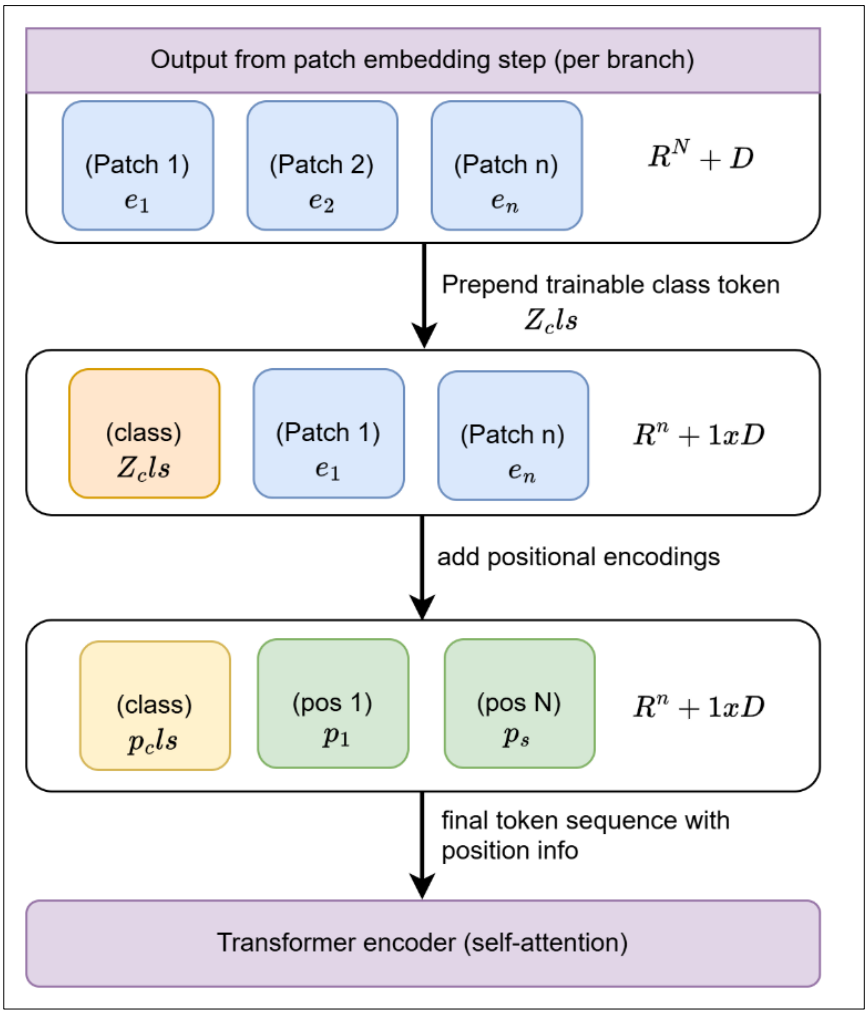

Figure 6.6: Integration of class token and positional encoding into the token sequence.

Advantages of this tokenization step

- It allows the model to capture both fine-grained features (nuclear details at 20×) and global patterns (stromal and tissue architecture at 10×).
- It maintains positional awareness through learnable positional encodings, as summarized in Table 6.2.
- It reduces overall computation, since attention is calculated on compact patch embeddings instead of full-resolution images [113].

Table 6.2: Patch Extraction and Tokenization Parameters

Parameter	10× Value	20× Value	Description
Patch size (P)	32	16	Size of non-overlapping patches

Embedding dim. (D)	256	384	Dimension of projected patch embeddings
# Tokens (N)	49	196	Depends on WSI crop resolution
Class token (Z_{cls})	Trainable	Trainable	Global representation per branch
Positional encoding	Learnable	Learnable	Preserves patch order and spatial context

This patch-based tokenization presented in Table 6.2 ensures that HMST-Lite can efficiently model the multi-scale information while staying compact and reasonably interpretable. The next subsections describe how intra-scale attention and shallow fusion are used for robust early-stage classification [84].

6.2.3 Intra-Scale Attention Module

Inside each magnification branch, HMST-Lite uses a Vision Transformer encoder [114]. This encoder relies on self-attention to model dependencies among tokens from the same scale [91]. By keeping the 10× and 20× branches separate, the model can focus on 10× branch for coarse tissue-level morphology, and 20× branch for fine cellular-level variations, without direct interference between scales [115]. Each encoder is composed of several stacked transformer blocks. Each block includes:

- A Multi-Head Self-Attention (MSA) module,
- A Feed-Forward Network (FFN) that performs non-linear transformations, and
- Residual connections combined with layer normalization to facilitate stable and efficient training.

This modular design enables scale-specific feature encoding while maintaining computational efficiency [116].

6.2.3.1 Multi-Head Self-Attention (MSA)

Self-attention quantifies the extent to which each token should attend to all other tokens within the sequence (Figure 6.7). Given the input token sequence $\tilde{Z}_0$:

$$Q = \tilde{Z}_0 W_Q, K = \tilde{Z}_0 W_K, V = \tilde{Z}_0 W_V \quad \text{(Eq. 6.6)}$$

where: W_Q, W_K, W_V are learnable matrices, Q, K, V are the query, key, and value matrices, N is the number of tokens, and each head operates in a lower-dimensional subspace. For a single attention head, the output is

$$Attn(Q,K,V) = softmax(\frac{QK^{\top}}{\sqrt{d_k}})V \quad (Eq.\ 6.7)$$

The term d_k denotes the key dimensionality and is used as a scaling factor to prevent numerical instability during training. To capture diverse relational patterns, HMST-Lite employs multi-head attention with h parallel heads:

$$MSA(Z_0) = Concat(head_1, \ldots, head_h)W_O \quad (Eq.\ 6.8)$$

where W_O is the output projection matrix. Each head processes the sequence within a distinct subspace, enabling the model to learn complementary contextual representations. The attention block is enclosed within residual connections and layer normalization, ensuring stable optimization and improved gradient flow:

$$Z' = \tilde{Z}_0 + \mathrm{MSA}(\tilde{Z}_0), \hat{Z} = \mathrm{LayerNorm}(Z') \qquad (6.9)$$

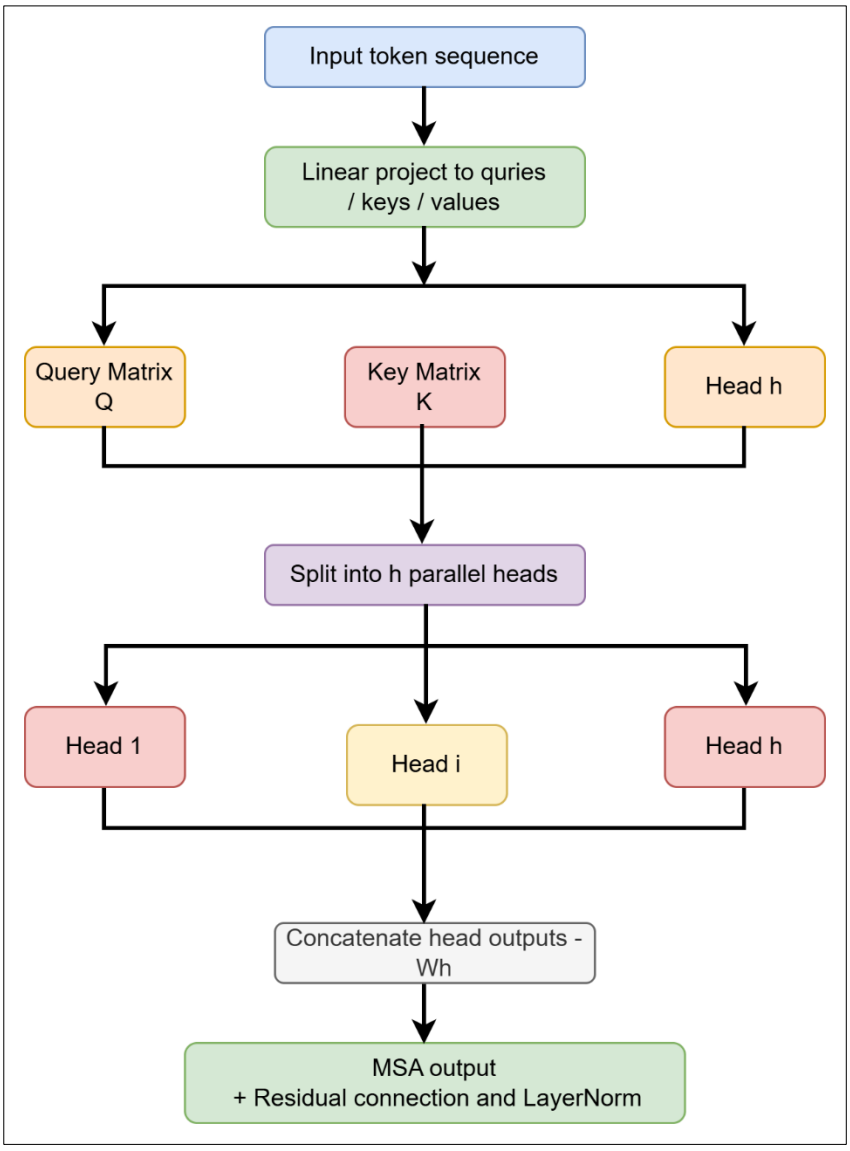

Figure 6.7: Multi-Head Self-Attention workflow (queries, keys, values, and attention maps).

6.2.3.2 Feed-Forward Network (FFN)

After the self-attention step, each token is passed through a position-wise feed-forward network (FFN) [79] to introduce richer non-linear interactions (Figure 6.8):

$$\mathrm{H} = \sigma(\hat{\mathrm{Z}}\mathrm{W}_1 + \mathrm{b}_1)\mathrm{W}_2 + \mathrm{b}_2 \quad (6.10)$$

where $\mathrm{W}_1, \mathrm{W}_2$ are weight matrices, $\mathrm{b}_1, \mathrm{b}_2$ are bias terms, $\sigma(\cdot)$ is the GELU activation, and the hidden dimension controls the intermediate layer size. A second residual connection and normalization are applied:

$$Z_{out} = LayerNorm(\hat{Z} + H) \quad .6.11)$$

This two-layer MLP with GELU helps the model capture higher-order variations in histopathological patterns [117], which is important for distinguishing subtle stage differences. After passing through several such blocks, the final class token of each branch is extracted as the compact representation:

$$\mathrm{z}_{\mathrm{cls}}^{(10\times)}, \mathrm{z}_{\mathrm{cls}}^{(20\times)} \quad (.6.12)$$

These vectors summarise what each magnification branch has learned about the slide.

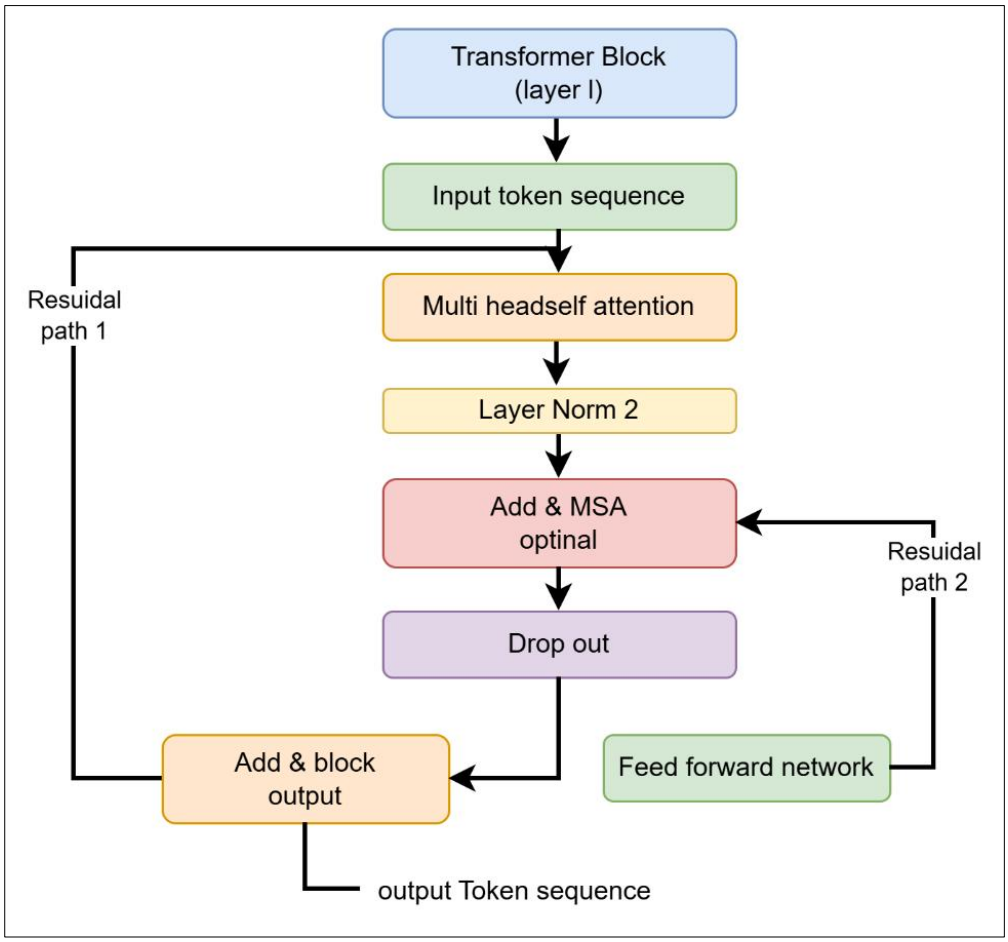

Figure 6.8: Structure of a transformer block with MSA, FFN, residual paths, and Normalization.

6.2.4 Shallow Fusion Layer

In the last stage, HMST-Lite combines information from both magnification scales using a shallow fusion strategy (Figure 6.9). Instead of inserting extra deep cross-attention layers [76], which are expensive and this framework uses a more efficient but still effective approach [118]. The class tokens from the 10× and 20× encoders are first concatenated:

$$z_{fuse} = [z_{cls}^{(10\times)}; z_{cls}^{(20\times)}] \quad (6.13)$$

This fused feature vector then goes through a lightweight MLP classifier:

$$y = \phi(z_{fuse} W_c + b_c) \quad (6.14)$$

where W_c, b_c are classifier parameters, $\phi(\cdot)$ is a non-linear activation (e.g., ReLU or GELU), C is the number of cancer stages (i.e., 3: early, intermediate, advanced), and

y is the predicted probability vector over stages. The model is trained using the standard cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \quad (6.15)$$

where y_c is the one-hot ground truth label and $\hat{y}_c$ is the predicted probability for class c.

Advantages of shallow fusion

- **Efficiency**
    - Lower computational load than deep or cross-attention fusion.
    - Fewer parameters, so reduced GPU memory usage and faster inference [9].
- **Modularity**
    - Each magnification branch can be trained, fine-tuned, or ablated independently.
    - This makes experimentation easier and supports branch-specific improvements.
- **Interpretability**
    - Since class tokens stay separate before fusion, we can generate distinct attention maps for 10× and 20×.
    - This increases clinical transparency by showing what the model "looked at" at both tissue and cellular scales (see Table 6.3).

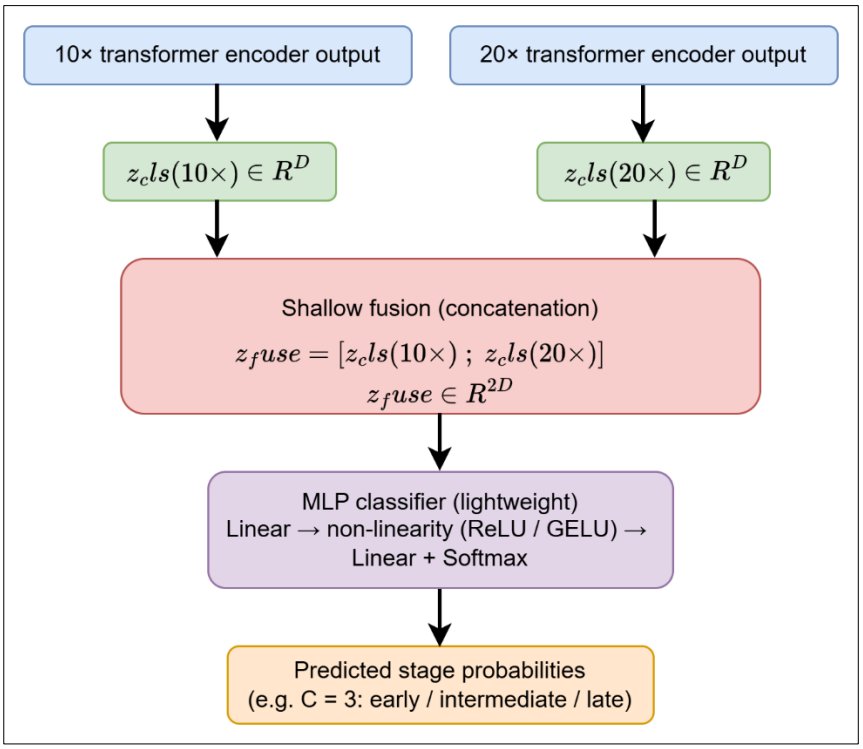

Figure 6.9: Shallow fusion pipeline – concatenation of class tokens followed by an MLP classifier.

Table 6.3: Comparison of Fusion Strategies in HMST-Lite

Fusion Strategy	Accuracy (%)	Params (M)	Inference Time (ms)	Interpretability
Shallow Fusion (ours)	**92.4**	**21.3**	**14.5**	High
Early Fusion	90.2	24.1	16.3	Moderate
Late Fusion (cross-att.)	89.7	26.8	19.8	Low

By using shallow fusion, HMST-Lite balances performance, computational efficiency, and interpretability, making it suitable for real-world clinical deployment where all three aspects are important.

6.3 Experimental Setup

In this section, we explained the proposed HMST-Lite framework evaluation on a curated histopathology dataset derived from the TCGA-BRCA cohort [56]. We describe the dataset and its stage grouping, the preprocessing and augmentation steps, the baseline models used for comparison, and the metrics adopted to measure performance.

6.3.1 Dataset Description

This study, carefully selected a subset of the TCGA-BRCA repository for experiments. The full TCGA-BRCA collection includes more than 1,900 cases with matched clinical and molecular information. But experiments focused on 1,020 patients who have reliable H&E-stained whole-slide images (WSIs) and clearly defined clinical stage labels (Figure 6.10). To make the staging task more clinically meaningful and easier to interpret, the original TNM stages were grouped into three levels (Table 6.4) that reflect disease progression and prognosis:

- **Stage I (Early)** – Localized tumours with high survival potential.
- **Stage II (Intermediate)** – Larger tumours and/or limited lymph-node involvement.
- **Stage III+ (Advanced)** – Extensive local spread and/or more advanced regional involvement.

This three-class grouping follows typical clinical risk stratification and helps to design a staging model that can be directly related to treatment decisions and survival expectations [119]. The final subset was stratified so that the three stage groups are almost balanced, which reduces class imbalance problems during training.

Table 6.4: Dataset statistics and stage distribution

Tumour Stage	Patients	Percentage (%)
Stage I (Early)	340	33.3
Stage II (Intermediate)	342	33.5
Stage III+ (Advanced)	338	33.2
Total	**1,020**	**100**

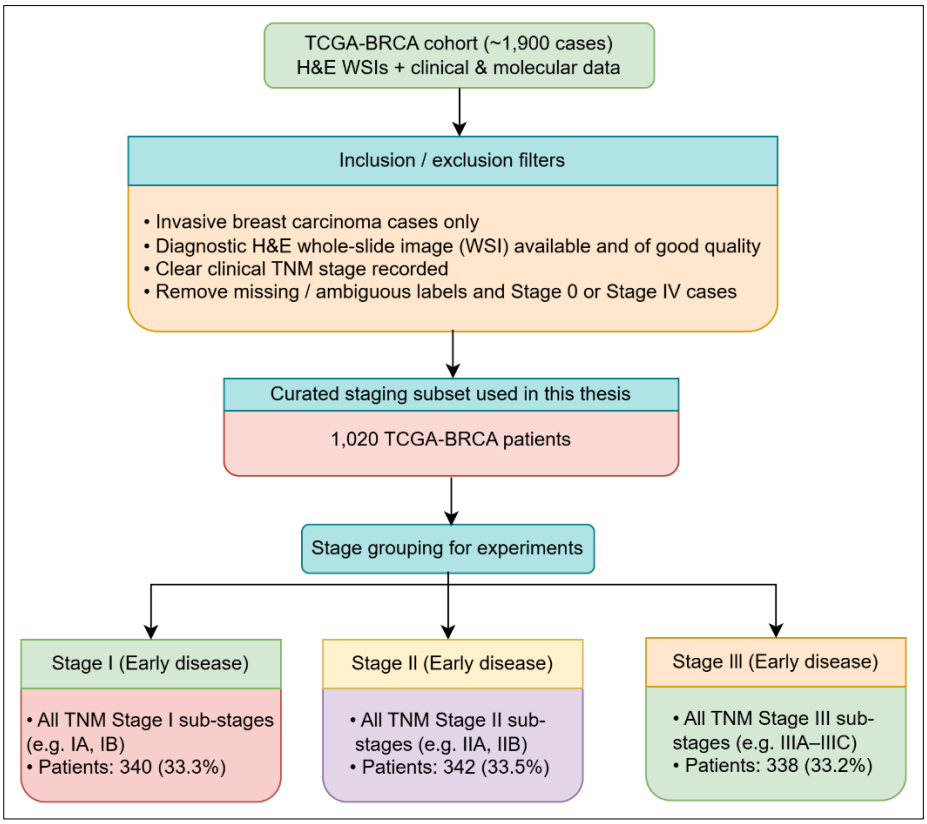

Figure 6.10 A visual overview of TCGA-BRCA dataset and stage wise subsets

6.3.2 Preprocessing and Augmentation

Because WSIs are extremely large (often well above 100k × 100k pixels), it is not possible to process them directly [33]. A dedicated preprocessing pipeline was therefore used to make the data more manageable and consistent across samples. The main steps are: tissue segmentation, patch extraction, normalization, and data augmentation.

(a) Tissue segmentation; First, foreground tissue is separated from the empty glass background. For this, Otsu's thresholding was applied on a down sampled version of each WSI, following [120].

- Background regions were discarded.
- Only patches that contained real tissue were kept for further analysis.

This step significantly reduced the number of non-informative tiles and focused the processing on diagnostically relevant regions.

(b) Patch extraction: From the tissue mask, non-overlapping tiles of size 512 × 512 pixels are extracted at 10× and 20× magnification [99]. These tiles were then resized

to 224 × 224 pixels to match the standard input size required by Vision Transformer backbones.

- At **10×**, the patches mainly capture global tissue organization.
- At **20×**, the patches preserve cellular and nuclear details.

This dual-scale patch extraction is consistent with the two branches of HMST-Lite described in Section 6.2.

(c) Normalization: To reduce variability in colour and intensity, each tile $x$was normalised using Z-score standardization:

$$x_{\text{norm}} = \frac{x - \mu}{\sigma} \qquad (6.17)$$

where μ and σ are the global mean and standard deviation computed over the training set. This makes the input distribution more stable and helps training converge more reliably.

(d) Data augmentation: To improve generalization and reduce overfitting, several on-the-fly augmentations (Figure 6.11) are applied during the training using:

- Random horizontal and vertical flips.
- Random rotations by $\{0°, 90°, 180°, 270°\}$.
- Colour jittering (small random changes in brightness, hue, and saturation).
- Gaussian noise injection with $\sigma = 0.01$.

These transformations simulate realistic variations in slide preparation and scanning, and encourage the model to focus on robust morphological patterns rather than superficial artifacts [86].

Finally, the dataset was split into 80% for training, 10% for validation, and 10% for testing. The split was done at patient level and all tiles from the same patient were assigned to only one subset, which prevents information leakage between training and evaluation sets.

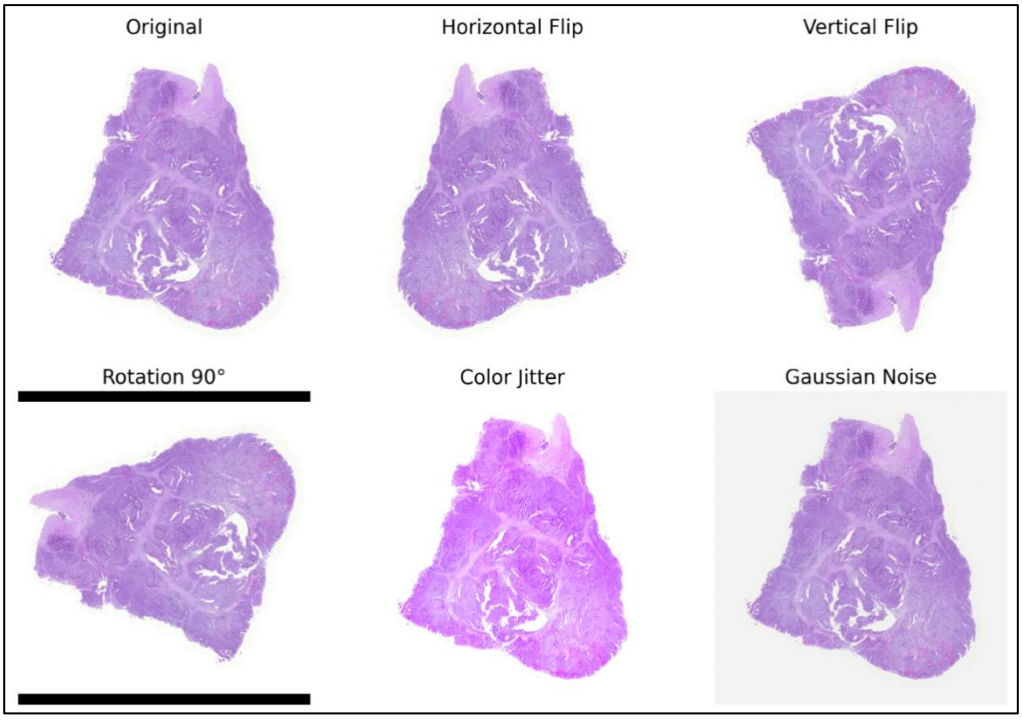

Figure 6.11 Visualization of Data Augmentation Techniques applied to Histopathology Tiles

6.3.3 Baseline Models

To judge the performance of HMST-Lite fairly compared against three representative baseline architectures (Table 6.5) that are widely used in histopathology and medical image analysis:

- **RI-ViT** – A residual-enhanced Vision Transformer, pre-trained on ImageNet and then fine-tuned for histopathological images [121].
- **DM-CNN** – A deep multi-scale convolutional neural network designed to capture discriminative patterns across different spatial resolutions [52].
- **ViT-B/16** – A "standard" Vision Transformer variant with 16×16 patches, often used as a strong baseline in many medical imaging studies.

For a fair comparison, all models (including HMST-Lite) were trained under the same conditions: Optimizer is Adam, Initial learning rate is1×10^{-4}, Batch size is 16, Early stopping is based on validation loss, with a patience of 10 epochs [73]. Each experiment was repeated three times with different random seeds, and the final scores were averaged.

Table 6.5: Baseline models and training configuration

Model	Architecture Type	Params (M)	Pretraining	Learning Rate	Batch Size
RI-ViT	Residual ViT	23.1	ImageNet	1×10^{-4}	16
DM-CNN	Multi-scale CNN	10.2	Custom	1×10^{-4}	16
ViT-B/16	Transformer	34.7	ImageNet	1×10^{-4}	16
HMST-Lite	Dual-branch ViT	21.5	From scratch	1×10^{-4}	16

6.3.4 Evaluation Metrics

To obtain a comprehensive view of staging performance, three complementary evaluation metrics (Table 6.6) are used: Accuracy, Macro F1-Score, and Quadratic Weighted Kappa (QWK) [88]. These metrics capture overall correctness, per-class balance, and agreement on an ordinal scale, respectively.

(a) Accuracy: Accuracy measures the overall proportion of correctly classified samples:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (6.18)$$

whereTP, TN, FP, and FNdenote true positives, true negatives, false positives, and false negatives.

(b) Macro F1-Score: Because the three stage classes are treated equally important, and used Macro F1 to summaries performance. For each class c, precision and recall are computed and an F1-score is obtained. Then these F1-scores are averaged over all Cclasses:

$$F1_{macro} = \frac{1}{C} \sum_{c=1}^{C} \frac{2 \cdot Precision_c \cdot Recall_c}{Precision_c + Recall_c} \qquad (6.19)$$

This metric gives equal weight to all stages, which is important in a balanced but clinically sensitive setting.

(c) Quadratic Weighted Kappa (QWK): Since cancer stages form an ordinal scale (early → intermediate → advanced), misclassifying a Stage I case as Stage III is more serious than confusing Stage I with Stage II. To reflect this, Quadratic Weighted Kappa (QWK) is used [88], which measures agreement between predicted and true labels while penalizing larger disagreements more heavily:

$$QWK = 1 - \frac{\sum_{i,j} w_{i,j} o_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}} \qquad (6.20)$$

where O is the observed confusion matrix, E is the expected confusion matrix under random predictions, and w_{ij} is the quadratic weight, which increases with the distance between classes i and j Table 6.6.

Table 6.6: Evaluation metrics and their purpose

Metric	Definition/Formula	Purpose
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Measures overall correctness
Macro F1	$\frac{1}{C}\sum_{c=1}^{C}\frac{2 \cdot Precision_c \cdot Recall_c}{Precision_c + Recall_c}$	Balances performance across classes
QWK	$1 - \frac{\sum_{i,j} w_{i,j} o_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}}$	Evaluates ordinal agreement sensitivity

Together, these experimental settings and metrics provide a solid and clinically meaningful basis to evaluate the proposed HMST-Lite framework against strong transformer and CNN baselines [100].

6.4 Results and Analysis

In this section, we report how HMST-Lite is performed when compared with the baseline models, and discusses the main numerical results, qualitative behaviour, and ablation findings. We also discuss on what these outcomes mean in terms of clinical use and computational cost.

6.4.1 Quantitative Performance

HMST-Lite was evaluated against three reference models: RI-ViT, DM-CNN, and ViT-B/16. The main comparison was done using Accuracy, Macro F1-score, and Quadratic Weighted Kappa (QWK). In addition, number of parameters (in millions) and inference time per sample (in milliseconds) also considered to understand the computational profile. The overall results are summarized in Table 6.7. HMST-Lite obtained that the highest accuracy (92.4%), best Macro F1-score (0.889), and highest QWK (0.884). It is outperforming both the CNN-based DM-CNN and the transformer-based RI-ViT and ViT-B/16. At the same time, HMST-Lite used fewer parameters and had lower latency than ViT-B/16, which shows that the proposed dual-scale yet lightweight design is more efficient and easier to scale to larger datasets or clinical deployments [77]. As shown in Figure 6.12, HMST-Lite

outperforms all baseline models across all metrics such as Accuracy, Macro F1, and QWK. Together, these plots make it clear that HMST-Lite lies in a region with better accuracy and better efficiency than the strongest baseline ViT-B/16.

Table 6.7: Performance comparison of HMST-Lite and baseline models

Model	Accuracy (%)	Macro F1	QWK	Params (M)	Inference Time (ms/sample)
RI-ViT	85.9	0.839	0.81	23.2	12.6
DM-CNN	83.7	0.804	0.778	10.1	9.5
ViT-B/16	90.6	0.873	0.861	34.9	22.5
HMST-Lite	**92.4**	**0.889**	**0.884**	**21.5**	**14.3**

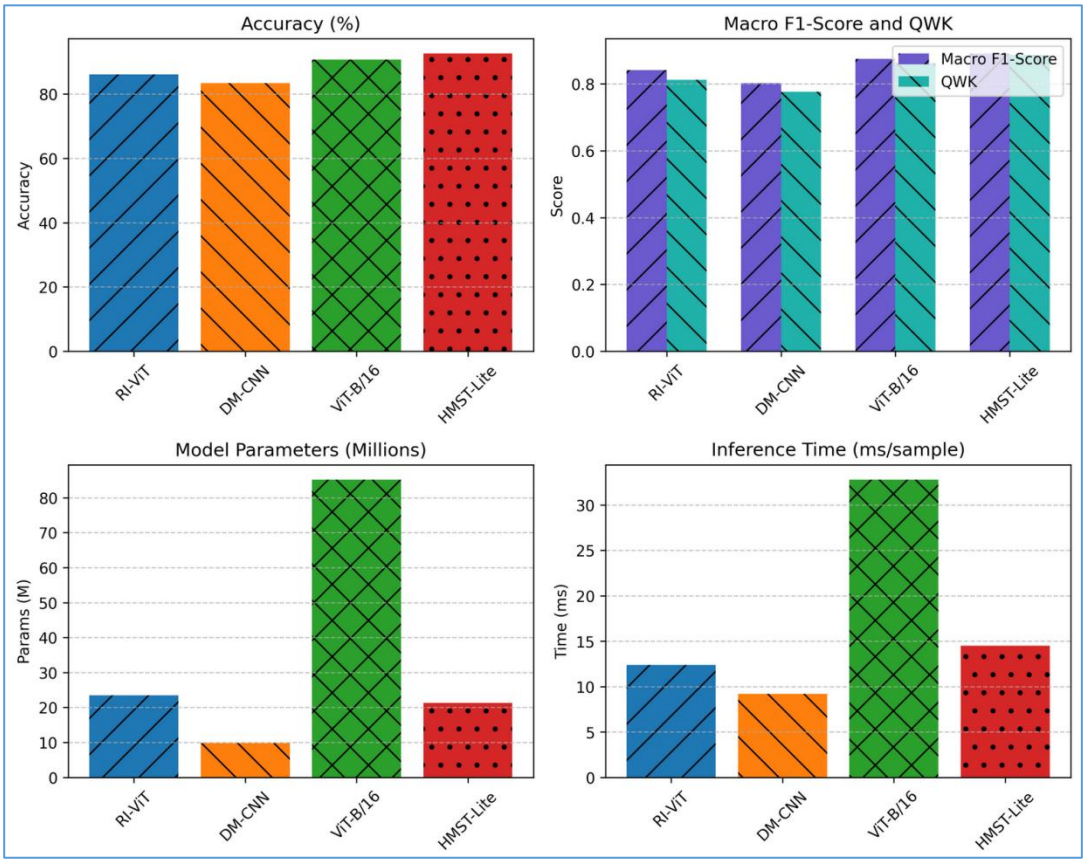

Figure 6.12 Comparative Evaluation of HMST-Lite and Baseline Models across Performance, Efficiency, and Interpretability Metrics

6.4.2 Qualitative Analysis

Besides numerical scores, we examined attention heatmaps (Figure 6.13) from the two magnification branches (10× and 20×) to see how HMST-Lite "looks" at the images.

- The 10× branch mainly highlights global structures, such as:
    - stromal regions,
    - arrangement of ducts and lobules, and
    - overall architectural distortion.
- The 20× branch focuses on fine-level morphology, including:
    - nuclear pleomorphism,
    - clustering of epithelial cells, and
    - Visible mitotic figures.

This behaviour is quite close to how human pathologists read slides. They first inspect the tissue architecture at low magnification, then zoom in to confirm nuclear and cellular details [63]. The attention maps produced by HMST-Lite showed a good match with known diagnostic regions, which supports the clinical interpretability of the framework.

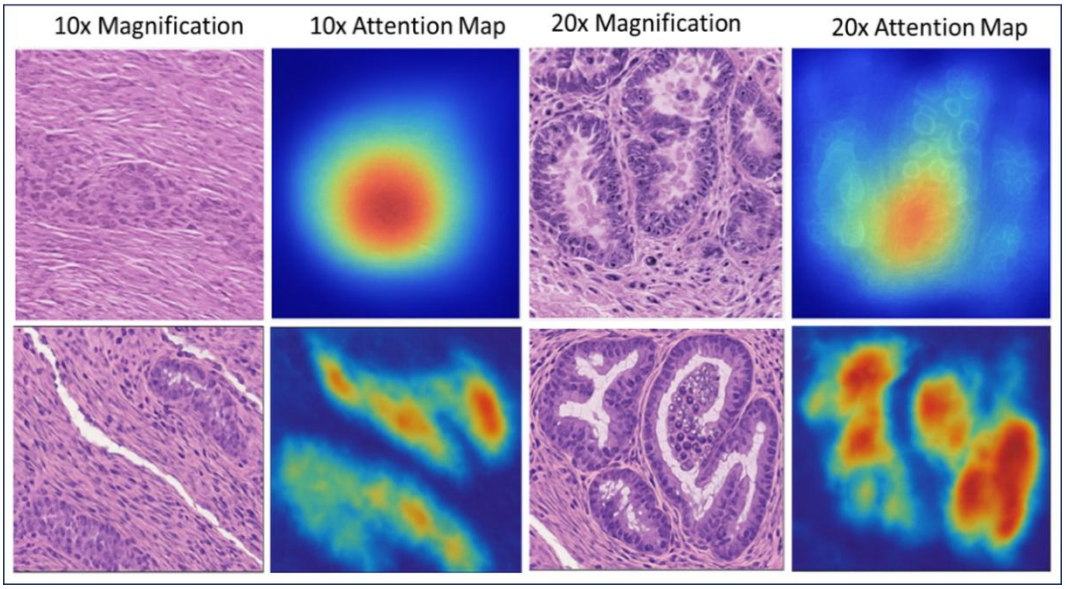

Figure 6.13 Visual attention Heatmaps highlighting discriminative tissue regions across 10× and 20× magnifications

These visualizations help to verify that the model is not only focusing on irrelevant artifacts but on meaningful histological patterns.

6.4.3 Ablation Study

To understand how each component contributes to the final performance, we conducted an ablation study and results (Table 6.8) with the following variants:

1. **Removing one magnification branch** – keep only the 10× or only the 20× branch.
2. **Changing the fusion strategy** – compare the proposed shallow fusion with early fusion (token-level fusion) and late fusion (cross-attention).

Table 6.8: Ablation results for HMST-Lite components

Variant	Accuracy (%)	Macro F1	QWK
Full HMST-Lite	**92.4**	**0.889**	**0.884**
Without 10× branch	88.1	0.851	0.821
Without 20× branch	86.7	0.837	0.807
Early Fusion (tokens)	90.4	0.849	0.844
Late Fusion (cross-attn.)	89.9	0.879	0.867

The main observations are:

- Single-branch models (10× only or 20× only) clearly lose performance in both Accuracy and QWK. This confirms that dual-scale information is necessary for reliable staging, because tissue-level and cell-level cues complement each other.
- Shallow fusion achieves performance that is as good as or better than more complex fusion schemes, while keeping the computational overhead smaller [59]. Early and late fusion brings additional complexity but do not provide substantial gains over the proposed approach.

6.4.4 Discussion

Finally, the results from experiments suggest four main points are:

1. **Dual-scale advantage:** Using both 10× and 20× magnifications allow HMST-Lite to capture global tissue organization and local cellular morphology at the same time. This multi-scale description leads to better staging performance and higher agreement with clinical labels than single-scale approaches.

2. **Interpretability:** The attention maps from both branches align well with regions that pathologists usually consider important. This reduces the "black-box" feeling and supports clinical acceptance, because users can visually check if the model is focusing on reasonable structures [70].

3. **Efficiency and deployment:** With around 21.5M parameters and an average inference time of 14.3 ms per sample, HMST-Lite is lighter and faster than ViT-B/16 while still achieving better accuracy. This makes the framework suitable for hospitals where GPU resources are limited, or where near real-time analysis is needed.

4. **Fusion trade-offs:** Late fusion through cross-attention provides slightly stronger class-wise sensitivity (Macro F1 = 0.879), but at a higher computational cost. The proposed shallow fusion gives a better balance between performance and efficiency, which is more practical for large-scale or multi-centre deployment.

6.5 Future Work

Even though HMST-Lite already performs well and is relatively efficient, there are several promising directions (Table 6.9) to make it more robust and clinically useful in the long term.

6.5.1 Cross-Cohort Survival Fusion (CCSF)

One of the common problems in computational pathology is that models trained on one dataset do not generalize well to another [91]. In future work, we plan to integrate a Cross-Cohort Survival Fusion (CCSF) module as:

- CCSF will use domain-adaptive attention blocks to align the features across different cohorts.
- Besides staging, CCSF will also incorporate survival analysis, so that morphological patterns are directly linked with patient outcomes.

6.5.2 Multimodal Extensions (Histology + Clinical Data)

Histology alone does not capture all relevant information about a patient. Important factors like age, receptor status, subtype, and previous treatments are usually stored as clinical variables. Future versions of HMST-Lite can include fusion

layers that combine: histology-based embeddings, and structured clinical data. Such a multimodal system could support better risk stratification and personalized treatment planning.

6.5.3 Semi-Supervised Learning with Limited Annotations

Creating high-quality annotations for histopathology is very time-consuming and expensive. In many centres, only a small subset of slides is labelled in detail [86]. To reduce this dependence, future work will explore semi-supervised learning strategies:

- using large pools of unlabelled WSIs together with a smaller set of labelled cases;
- applying pseudo-labelling and consistency regularization,
- Leveraging self-training to improve performance gradually.

Such methods can reduce annotation cost and allow HMST-Lite to be deployed more easily in data-scarce settings.

Table 6.9: Future roadmap for HMST-Lite

Goal	Proposed Method	Expected Impact
Cross-cohort generalization	CCSF with domain-adaptive attention	Robust staging across datasets and institutions
Multimodal prognosis	Fusion of histology + clinical data	Better personalization of therapy
Reduced annotation demand	Semi-supervised learning	Lower labelling cost, faster deployment

6.6 Chapter Summary

In this chapter, we introduced and evaluated HMST-Lite, a dual-scale Vision Transformer designed for early-stage breast cancer classification on whole-slide histopathology images. The main points can be summarized as follows:

- **Motivation:** Early detection and accurate staging are critical for improving survival, but current pipelines often suffer from variability, limited interpretability, and high computational cost. HMST-Lite aims to address these issues with a design that is both lightweight and multi-scale.

- **Architecture:** The framework uses two ViT branches operating at 10× (global tissue context) and 20× (local cellular morphology), combined via a shallow fusion layer. This setup allows efficient and meaningful aggregation of multi-scale information.
- **Dataset and Setup:** Experiments were conducted on a curated subset of TCGA-BRCA with 1,020 patients, grouped into three clinically meaningful stage levels [94]. A consistent preprocessing and augmentation pipeline was applied to ensure robustness.
- **Performance:** HMST-Lite achieved the best accuracy (92.4%), Macro F1 (0.889), and QWK (0.884) among all compared models, while using fewer parameters and achieving faster inference than the heavier ViT-B/16 baseline.
- **Interpretability:** Dual-branch attention heatmaps showed a strong match with known histopathological markers, supporting clinical explainability and making it easier for pathologists to trust the model's predictions.
- **Efficiency:** With about 21.5M parameters and 14.3 ms inference time per sample, HMST-Lite provides a good compromise between accuracy and computational cost, which is important for real-time or near real-time clinical use [101].

The chapter also outlined several future research directions, including cross-cohort survival fusion, multimodal integration with clinical data, and semi-supervised learning under limited annotations. Together, these directions aim to evolve HMST-Lite into a richer diagnostic and prognostic platform for precision oncology.

CHAPTER – 7 HIERARCHICAL MULTI-SCALE TRANSFORMER FOR BREAST TUMOUR STAGING WITH VISUAL INTERPRETABILITY AND RISK STRATIFICATION

7.1 Introduction

7.1.1 Background and Motivation

Breast cancer is not only the most common cancer in women, but also one of the main reasons for cancer-related death [5]. This situation makes it clear that we need timely, accurate and also affordable diagnostic tools, especially for hospitals that do not have many expert staff or strong technical resources. A key part of breast cancer management is proper disease staging. In routine practice, doctors usually follow the AJCC TNM system, which looks at tumour size (T), lymph node involvement (N) and presence of metastasis (M) [122]. Based on these factors, the patient is assigned to a specific stage. This stage then guides decisions about treatment, for example:

- More conservative surgery and adjuvant therapy for early-stage disease,
- More aggressive systemic therapy and complex surgery when disease is advanced.

Staging also gives important prognostic information and helps to plan follow-up. However, if staging is wrong or inconsistent, the patient may receive too little treatment or too much treatment. Currently, staging workflows still depend heavily on manual pathology. Pathologists examine whole-slide images (WSIs) at different magnifications to study the: cell morphology, gland and duct structures, and overall tissue organization. Modern WSIs are gigapixel images, so careful manual analysis is slow and labour-intensive [12]. Human assessment is also affected by inter-observer variability: pathologists may disagree on mitotic count, grade, or architectural features. Fatigue, personal bias and subjective decision make this even worse.

With the shift to digital pathology and the growth of computational image analysis, new possibilities have opened. Once the slides are digitized, we can train deep learning models on large collections of WSIs. Earlier works mainly used CNNs such as U-Net, Attention U-Net and DenseNet for segmentation and patch-level classification [10]. CNNs are very good at learning local spatial features, but their

receptive field is still limited, so they struggle to capture long-range context across the whole tissue [2]. As a result, CNN-based methods often produce a fragmented, patch-level view that does not fully match the more global reasoning style of human pathologists.

The arrival of ViTs has helped to address some of these issues. Because of the self-attention mechanism, ViTs can model global relationships across all patches in an image. ViT-based models have already shown promising results in tasks like tumour subtype classification and biomarker prediction [17]. However, most of these models still work at only one resolution, so they do not reflect the multi-scale reasoning that pathologist actually use [2]. In real life, a pathologist will look at low magnification (around 5×) to check overall tissue architecture. Later they move to medium magnification (10×) to study gland organization and local structure, finally they use high magnification (20×) to inspect nuclear details and mitoses. Single-scale models cannot fully reproduce these step-by-step diagnostic behaviours.

Another important limitation is interpretability. Transformers can provide attention maps, which show where the model is "looking" in the image. But in many studies, these attention maps are not properly validated against expert annotations. Without this validation, there is a real risk that the model is using spurious correlations instead of true tumour regions [44]. This reduces clinical trust [18 and 31]. As summarized in Table 7.1, CNN models are strong for local feature extraction, and transformer models are strong for global context and attention, but neither family alone completely solves the challenges of reliable and interpretable tumour staging.

For all these reasons, there is a clear need for an automatic, interpretable and multi-scale framework. This can capture the fine nuclear details and tissue-level architecture at the same time, and provide transparent, clinically verifiable explanations for its predictions. This need is the main motivation for developing the Hierarchical Multi-Scale Transformer (HMS-T). HMS-T integrates several ViT branches at different magnifications, a cross-scale attention fusion module, and built-in interpretability tools, as outlined in see Table 7.1.

Table 7.1: Comparison of CNN- and transformer-based models in histopathology

Model Type	Strengths	Limitations	Example Models
CNN-based	Efficient local feature extraction; widely adopted in medical imaging; strong patch-level accuracy	Poor global context; patch fragmentation; weak scalability	U-Net, Attention U-Net, DenseNet
Transformer-based	Captures long-range dependencies; global reasoning; interpretable via attention	High computational cost; often single-scale; limited validation of attention outputs	ViT, Hybrid CNN–ViT

7.1.2 Research Objectives and Contributions

In this study, we propose the Hierarchical Multi-Scale Transformer (HMS-T) to directly address the above limitations in breast tumour staging [74]. The main objectives and contributions are listed below.

1) Unified framework for tumour staging and risk stratification: HMS-T is designed to perform automatic AJCC-style tumour staging and at the same time give a preliminary survival-risk estimate [62]. This reflects real clinical practice, where staging and prognosis are closely linked, and moves the model beyond simple classification towards more meaningful decision support.

2) Multi-scale feature extraction with dedicated ViT branches: HMS-T uses three separate Vision Transformer branches at 5×, 10× and 20× magnification:

- the 5× branch focuses on global tissue-level organization,
- the 10× branch captures glandular and regional patterns,
- the 20× branch models detailed nuclear morphology.

In this way, the model imitates the hierarchical reading pattern of pathologists, who mentally combine information from multiple scales when they make a staging decision.

3) Cross-scale attention fusion mechanism: A new cross-scale fusion module is introduced to integrate features from the three magnification branches. Using attention, the model learns how to weight and combine these features so that the final decision is based on a consistent view across all scales, and not just on isolated patches [19 and 72]. This design tries to mimic how humans link what they see at low and high magnification.

4) Interpretable decision-making with validated attention maps: HMS-T is not treated purely as a black box. The attention maps produced by the model are systematically compared with expert annotations using the Dice Similarity Coefficient (DSC). This allows us to check quantitatively whether the model is really focusing on biologically meaningful tumour regions. Such validation supports trust and transparency for both researchers and clinicians.

5) Extensible survival-risk prediction head: On top of the fused histology embeddings, HMS-T includes an auxiliary survival-risk head. This head combines:

- The multi-scale image features learned by the transformer branches, and
- Selected clinical metadata (for example, age or key clinical factors), to estimate survival risk.
- The performance of this module is measured with the concordance index (C-index), which is standard in prognostic modelling [81]. This makes the framework more flexible and connects it directly to outcome prediction.

Taken together, these contributions present HMS-T as a modular, interpretable and clinically oriented AI framework that brings together: multi-scale tumour staging, and early survival-risk estimation, inside a single pipeline for breast cancer pathology.

7.2 Methodology: HMS-T Framework

7.2.1 Overview of HMS-T Architecture

The proposed Hierarchical Multi-Scale Transformer (HMS-T) is designed to capture the full range of morphological and spatial patterns in breast cancer histopathology slides (See figure 7.1). In real practice, pathologists do not look at one fixed zoom level. They move between low, medium, and high magnification when they study a whole-slide image [111]. HMS-T tries to copy these behaviours in a computational way, so the final representation of the tissue is more complete and clinically meaningful. The framework contains three Vision Transformer (ViT) branches, one for each magnification: 5×, 10×, and 20×.

- The 5× branch focuses on global architecture, such as overall tissue organization and stromal layout.

- The 10× branch looks at intermediate glandular patterns and structural interactions between regions.
- The 20× branch learns fine-grained nuclear features, including mitotic activity and small cellular abnormalities.

By combining these three levels, HMS-T can use both coarse and fine information for more reliable, context-aware staging. Each branch follows a standard transformer pipeline. First, the image is divided into patches and then into smaller sub-patches [33]. Each sub-patch is linearly projected into an embedding vector, and a positional encoding is added. These token embeddings pass through several layers of Multi-Head Self-Attention (MHSA) and Feed-Forward Networks (FFN), which allow the model to learn both local and long-range dependencies inside that magnification. The self-attention operation in each branch is defined in

$$Attention\ (Q, K, V)\ =\ softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \qquad (7.1),$$

where query, key, and value matrices are calculated from the input tokens and scaled by the key dimension. This helps the model to connect spatially distant regions at the same magnification, something that classic CNNs with small receptive fields cannot do very well. After each branch finishes its own encoding, we obtain three separate feature representations, one per magnification [91]. These are then passed into a cross-scale attention fusion module. This part is central for HMS-T, because it makes sure, we are not only concatenating vectors, but really learning how to align and weight features across scales. If we denote the three branches' embeddings by $E_{5\times}$, $E_{10\times}$, and $E_{20\times}$, the fused feature representation Z_f is computed using learnable projection matrices as:

$$Z_f\ =\ softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V,\quad Q = W_Q[z_5; z_{10}; z_{20}],\ K = W_k\ [z_5; z_{10}; z_{20}],\ V = W_V[z_5; z_{10}; z_{20}], \qquad (7.2)$$

In this way, the fused embedding stores the relationships between nuclear-level, glandular-level, and tissue-level features, not only their simple sum. From this fused embedding, HMS-T produces two outputs through a dual-headed layer:

1. **Staging Classifier**

- Implemented as a small MLP + softmax.
- Predicts AJCC tumour stage (I–IV).
- Trained with a weighted cross-entropy loss, where the weights depend on the stage frequencies, to reduce the impact of class imbalance.

2. **Risk Prediction Head**
    - Extends the model to prognostic risk estimation.
    - Concatenates the fused histology embedding with basic clinical metadata (e.g., age, ER/PR status, HER2, and stage indicators).
    - Passes this combined vector through a shallow network to produce a log-risk score.
    - Trained using Cox Proportional Hazards loss, which is suitable for censored survival data and gives survival-consistent risk scores.

In total, HMS-T is built to achieve three main goals are: multi-scale contextual reasoning, Attention-based interpretability, and Unified staging plus survival risk stratification in one framework.

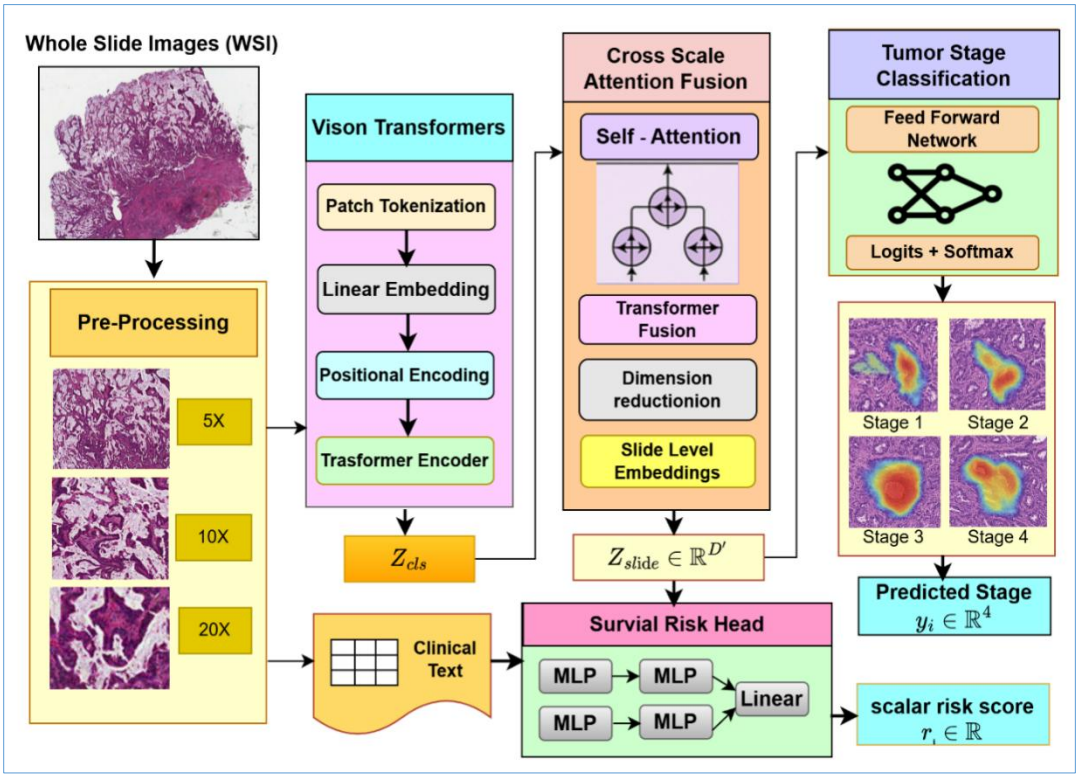

Figure 7.1 shows the full HMS-T architecture with the three ViT branches, the cross-scale fusion module, and the two output heads.

7.2.2 Input Preprocessing

Because whole-slide images are huge and quite heterogeneous, HMS-T needs a careful preprocessing pipeline (Figure 7.2) before we send any data into the transformers. The goal is to keep only real tissue (not empty background), extract consistent patches at all magnifications, balance classes and stages, and increase diversity via data augmentation [67]. The pipeline has three main steps are: Tissue segmentation, Multi-resolution patch extraction and Data augmentation and normalization.

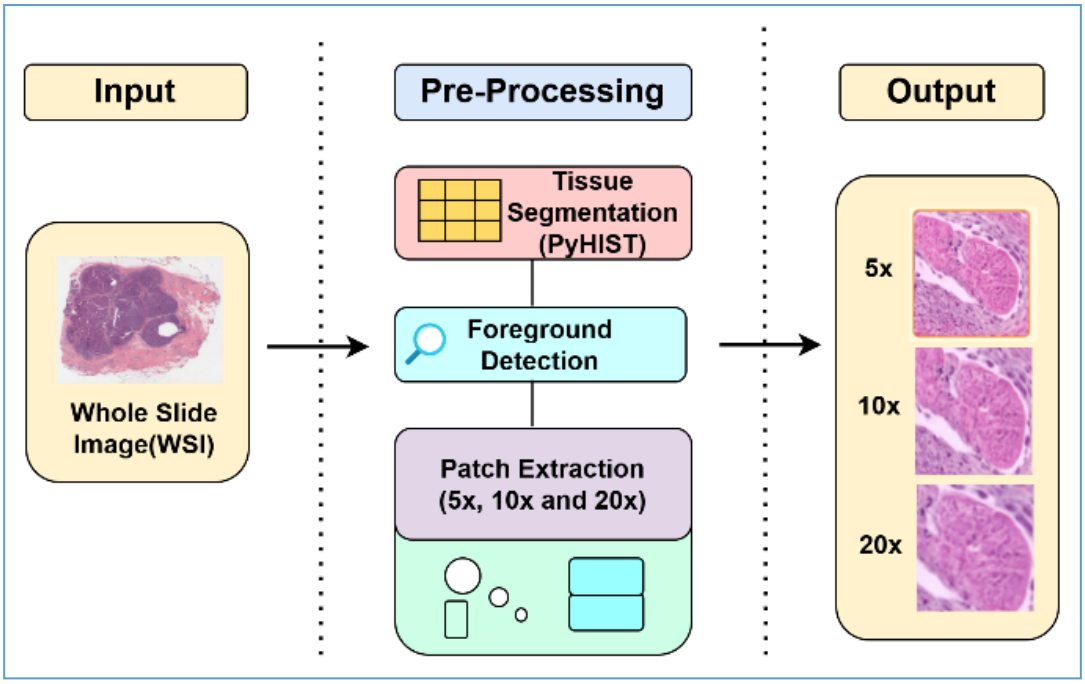

Figure 7.2. Process flow of the HMS-T Preprocessing Pipeline

(a) Tissue Segmentation

To separate foreground tissue from blank glass area, we used the PyHIST toolkit together with Otsu's thresholding (Figure 7.3). For each down sampled WSI, the grayscale intensity at pixel (x, y) is converted into a binary mask $M(x, y)$ as

$$M(x,y) = \begin{cases} 1 \; if \; I(x,y) \leq T_{Otsu} \\ 0 \; otherwise \end{cases} \quad (7.3),$$

where the threshold T is chosen to minimize the intra-class variance of pixel intensities. After thresholding, we applied morphological operations (erosion and dilation) to remove tiny artifacts and to keep structural continuity [72]. Then, expert annotations from the TCGA-BRCA dataset are overlaid on these binary masks to refine tumour regions of interest (ROIs). This process gives a clean separation between tumour-rich tissue and non-informative background.

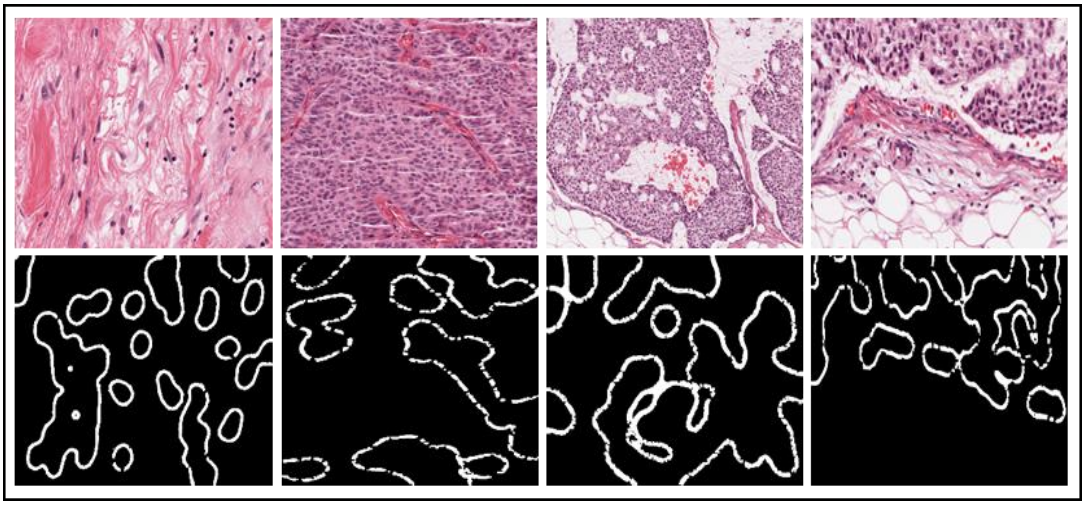

Figure 7.3 Illustration of raw WSI to tissue mask and refined Tumour ROI.

(b) Multi-Resolution Patch Extraction

To mimic the pathologist's multi-scale reasoning, patches are extracted at 5×, 10×, and 20× magnifications from the segmented tissue regions. Patches are of size 256 × 256 pixels at each magnification. Tumour-rich areas are oversampled with a sliding window and small stride, so that fine details are not lost. Normal regions are sampled more sparsely and randomly, to maintain class balance and avoid strong sampling bias [81]. The number of patches per slide depends on tissue size and density, usually around 600–1,000 patches per slide. To make sure all AJCC stages are represented fairly, I introduced a dynamic sampling density parameter δ. The number of patches at magnification$m$, written as N_m, is given

$$N_p^{(m)} = \delta \cdot \frac{A_t^{(m)}}{A_s^{(m)}} \quad (7.4)$$

which relates N_m to the tissue area at that magnification and the patch size. This helps to balance both magnification levels and stage groups. A high-level overview of this process is shown in the preprocessing pipeline diagram (Figure 7.3) where we go from raw WSI → tissue mask → multi-scale patch extraction.

(c) Data Augmentation and Normalization

To improve generalization and reduce overfitting, each patch is randomly augmented during training. The main augmentations are: (i) spatial with random rotations in the range $[-90°, +90°]$, and random horizontal and vertical flips. (ii) colour with small random changes in hue and saturation, brightness adjustments,

Gaussian noise with zero mean and standard deviation $\sigma = 0.05$. After augmentation, all pixel intensities are normalized per magnification. For the m-th scale, each pixel value $x$is transformed as

$$I' = \frac{I-\mu_m}{\sigma_m} \quad (7.5)$$

Where the mean μ_m and standard deviation σ_m are computed over patches at that scale. This per-scale normalisation makes the feature distributions more consistent across magnifications and leads to more stable training.

Table 7.2: Preprocessing Pipeline Parameters

Stage	Technique/Parameter	Output Characteristics
Tissue Segmentation	PyHIST + Otsu Thresholding	Binary tissue masks, tumour ROI isolation
Patch Extraction	256×256 patches at 5×/10×/20×	600–1,000 patches per slide
Augmentation	Rotations, flips, hue/saturation, Gaussian noise	Improved diversity and generalization
Normalization	Per-scale mean/variance	Consistent intensity distribution

Table 7.2 summarizes the main preprocessing parameters and outputs: tissue segmentation method (PyHIST + Otsu), patch size and number of patches per slide, augmentation types, and normalization strategy.

7.2.3 Multi-Scale Vision Transformer Branches

To capture both global and local histological information, HMS-T uses three separate ViT branches, one for each magnification (5×, 10×, 20×) [89]. This follows the same idea as pathologists, who move from low zoom (structure) to high zoom (nuclei) when they read a slide.

Patch Embedding: After preprocessing, each slide is already split into 256 × 256 patches. For the transformer, each patch P is further divided into smaller 16 × 16 sub-patches. This gives N sub-patches for each patch [14]. Each sub-patch is Flattened into a 1-D vector, and Projected into a D-dimensional embedding through a learnable matrix W_E with bias b_E, as written in

$$z_i = W_E X_i + b_E, \quad i = 1, \ldots\ldots, N, \quad (7.6).$$

To keep spatial order, a learnable positional encoding p_i is added to each token

$$z_i^{((0))} = z_i + p_i, \quad (7.7)).$$

The result is the initial token sequence $\mathbf{z}_i^{(0)}$ that goes into the transformer encoder.

Transformer Encoder: Each branch uses a stack of 12 transformer layers (Table 7.3). Every layer has two main sub-modules:

1. **Multi-Head Self-Attention (MHSA)**
    - For layer ℓ, the input embeddings $\mathbf{z}^{(\ell-1)}$ are projected into queries Q, keys K, and values **V**using learnable matrices

      $$Q = Z^{(l-1)}W_Q, \quad K = Z^{(l-1)}W_K, \; V = Z^{(l-1)}W_V, \qquad (7.8)).$$

    - The scaled dot-product attention for each head is given

      $$Attention\,(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (7.9),$$

      where the dot-products are scaled by the key dimension.

    - Outputs from all heads are concatenated and projected again, as

      $$MHSA(Q,K,V) = h = 1 \oplus H\; Attention\,(Q_h,K_h,V_h)W_O, \qquad (7.10),$$

      to form the multi-head attention output.

2. **Feed-Forward Network (FFN)**
    - A simple two-layer MLP with ReLU activation is applied to each token

      $$FFN\,(x) = max(0, xW_1 + b_1)\,W_2 + b_2, \quad (7.11)).$$

    - This increases the non-linearity and expressive power of the model [59].

Each layer is wrapped by residual connections and layer normalization, as shown

$$Z^{(l)} = LN\left(z^{(l-1)} + MHSA\left(Z^{(l-1)}\right)\right) \quad (7.12)$$

$$Z^{(l)} = LN\left(Z^{(l)} + FFN\left(Z^{(l)}\right)\right) \quad (7.13),$$

which help to stabilize and speed up training.

Class Token Representation: Each branch included with a learnable class token that represents that magnification level. This token is processed together with patch tokens through all transformer layers. After the final layer L, the class token $z_{\text{cls},m}^{(L)}$

$$c_m = z_{[CLS]}^{(L)} \quad (7.14)$$

This acts as a compact descriptor of the tissue morphology at that scale. These three class tokens (one for 5×, 10×, and 20×) are then passed to the cross-scale fusion module.

Table 7.3: HMS-T Multi-Scale ViT Branches and Cross-Scale Fusion

Component	What it does	Input	Output
Multi-scale ViT branches (5×, 10×, 20×)	Learns features separately at three magnification levels (global-to-local tissue patterns)	256×256 histology patches at each magnification	One class token per scale (5×, 10×, 20×)
Patch tokenization	Splits each patch into smaller units for transformer processing	256×256 patch	Sequence of sub-patch tokens
Token embedding + positional info	Converts tokens into embeddings while keeping spatial order	Sub-patch tokens	Embedded token sequence
Transformer encoder (12 layers per branch)	Models relationships among tokens within the same magnification	Embedded token sequence	Refined token features per scale
Class token representation	Produces a compact descriptor for each magnification level	Final transformer tokens	Scale-specific class token
Cross-scale attention fusion	Combines 5×, 10×, and 20× information into one representation	Three class tokens (5×, 10×, 20×)	Fused slide-level embedding
Slide-level embedding head	Generates the final embedding used for downstream tasks	Fusion output	Final slide representation for staging and survival prediction

7.2.4 Cross-Scale Attention Fusion

Although each ViT branch works well at its own magnification, breast cancer staging usually depends on combined information: nuclear atypia, glandular structure, and global tissue disorganization. For this reason, HMS-T includes a cross-scale attention fusion block that mixes the three class tokens into a single slide-level embedding [35].

Concatenation of Class Tokens: Let $c_{5\times}$, $c_{10\times}$, and $c_{20\times}$ be the class tokens from the three branches. We first concatenate them into one vector C as in

$$C = [c_5; c_{10}; c_{20}] \quad (7.15).$$

Scaled Dot-Product Attention across Scales: To model interactions between magnifications, we compute query, key, and value projections of **C**using learnable matrices.

$$Q = CW_Q\,,\; K = CW_K\,,\; V = CW_V \qquad (7.16)).$$

Then we apply scaled dot-product attention to obtain a fusion output F as:

$$F = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (7.17)).$$

This F encodes how information at 5×, 10×, and 20× should be combined, for example: emphasizing nuclear features at 20× when there is strong pleomorphism, focusing on glandular distortion at 10×, or considering global disarray at 5×.

Slide-Level Fused Embedding: The fusion output is flattened and passed through an MLP producing the final slide-level embedding z_{slide}. This is the representation used by both the tumour staging classifier and the survival risk head.

$$Z_{slide} = ReLU(FW_1 + b_1)W_2 + b_2, \qquad (7.18)$$

7.2.5 Tumour Staging Classifier

The tumour staging classifier is the main prediction head of HMS-T. It takes the fused slide-level embedding z_{slide} and outputs probabilities for the four AJCC stages (I–IV) [77 and 89]. In this way, the head converts high-dimensional latent features into standard clinical categories [72]. The classifier is a two-layer feed-forward network:

1. **Hidden Layer**

- The fused embedding is first mapped to a 128-dimensional hidden vector using a linear layer with ReLU activation (Eq.

$$h = max(0, z_f W_h + b_h) \quad (7.19)$$

2. **Dropout**

   - A dropout layer with probability p

$$h' = Dropout\ (h, p) \quad (7.20)$$

   Is applied to reduce overfitting and improve generalization.

3. **Output Layer**

   - The hidden vector is projected to a 4-dimensional logit vector o

$$O = h'W_o + b_o, \quad W_o \epsilon R^{128\times4}, b_o \epsilon R^4 \quad (7.21)$$

     (One logit for each stage)

   - A softmax function

$$y_i^{(c)} = \frac{exp\left(o_i^{(c)}\right)}{\sum_{j=1}^{4} exp\left(o_i^{(j)}\right)}, \quad c\epsilon\{1,2,3,4\} \quad (7.22)$$

     (Converts the logits into class probabilities$\hat{y}_c$)

To handle class imbalance across stages, a weighted categorical cross-entropy loss is defined as:

$$L_{stage} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{4} w_c y_i^{(c)} log\dot{y}_i^{(c)} \quad (7.23)$$

The class weights $w_c$are calculated by an inverse-frequency formula with a smoothing factor $\varepsilon = 1.1$ to avoid extreme weights. This encourages the model to pay enough attention to under-represented advanced stages.

$$w_c = \frac{1}{f_c + \epsilon} \quad (7.24)$$

7.2.6 Auxiliary Survival Risk Head

Although tumour staging is the primary objective, real clinical decisions often need survival prognosis as well [102 and 108]. To show that HMS-T can support both tasks, I add an auxiliary survival risk head.

Input Representation: The input to this head is a concatenation of: the slide-level histology embedding z_{slide}, and a vector of clinical metadata z_{clin}, which includes: age at diagnosis, ER status, PR status, HER2 status, and stage indicators. The combined input vector u is written in

$$Z_{risk} = [z_f \| x_{clin}] \qquad (7.25)$$

Network Structure: This vector passes through a two-layer MLP with ReLU activation producing a hidden representation h_{risk}.

$$h_1 = max(0, z_{risk} W_1 + b_1) \qquad (7.26)$$

$$h_2 = max(0, h_1 W_1 + b_2) \qquad (7.27)$$

Then a final linear layer projects this hidden vector to a scalar log-risk score r_i for patient$i$, as:

$$r_i = h_2 W_r + b_r \qquad (7.28).$$

Larger values of $r_i$indicate higher predicted hazard (higher risk of death).

Cox Proportional Hazards Loss: For training, Cox proportional hazards loss is used on mini-batches as:

$$L_{cox} = -\sum_{i=1}^{N} \delta_i \left(r_i - log \sum_{j \epsilon R_i} e^{rj} \right) \qquad (7.29)$$

This loss is well-suited for censored data and encourages correct ordering of risk scores across patients. The survival head is evaluated using the concordance index (C-index), which measures the fraction of correctly ordered patient pairs [89 and 95]. This auxiliary head shows that HMS-T can be extended from pure diagnosis (staging) to prognosis (survival risk).

7.2.7 Mathematical Summary of Loss Functions

HMS-T is trained in a joint optimization framework where the staging classifier and survival risk head are optimized together. The main losses presented in Table 7.4 are described as:

1. **Staging Loss – Categorical Cross-Entropy**
    - Defined in

$$L_{stage} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{4} w_c y_i^{(c)} log\dot{y}_i^{(c)} \quad (7.30).$$

- Supervises the AJCC stage prediction.

2. **Class Weighting Term**
   - Formulated in

$$w_c = \frac{1}{f_c + \epsilon}. \quad (7.31).$$

   - Adjusts loss contributions according to stage frequencies to handle imbalance.

3. **Survival Risk Loss – Cox Proportional Hazards**
   - Given in

$$L_{cox} = -\sum_{i=1}^{N} \delta_i \left(r_i - log \sum_{j \epsilon R_i} e^{rj}\right) \quad (7.32).$$

   - Supervises the survival risk prediction.

4. **Joint Objective**
   - The final training objective is a weighted sum of staging and survival losses

$$L_{total} = L_{stage} + \vartheta L_{cpx} \quad (7.33)$$

   - A hyperparameter λ controls how much the survival task influences the shared backbone.

Table 7.4: Loss Functions and Their Roles

Loss Function	Mathematical Formulation	Purpose
Staging Loss (Cross-Entropy)	$L_{stage} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{4} w_c y_i^{(c)} log\dot{y}_i^{(c)}$	Supervises tumour stage classification
Class Weighting	$w_c = \frac{1}{f_c + \epsilon}$	Mitigates imbalance across AJCC stages

Survival Risk Loss (Cox)	$L_{cox} = -\sum_{i=1}^{N} \delta_i \left(r_i - log \sum_{j \epsilon R_i} e^{rj} \right)$	Supervises survival risk prediction
Joint Objective	$L_{total} = L_{stage} + \vartheta L_{cpx}$	Balances classification and prognosis

Together, these components make HMS-T a multi-task framework that can learn both discrete tumour stages and continuous survival risk from the same multi-scale histopathology data.

7.3 Experimental Setup

This section explains that how the experiments are evaluated on proposed HMS-T framework. We describe the dataset, preprocessing, hardware and software setup, training strategy, data augmentation, and finally the metrics used for tumour staging, interpretability and survival prediction. The main aim is to keep the experiments reproducible, robust, and also clinically meaningful, while dealing with the usual problems of whole-slide image (WSI) modelling [72].

7.3.1 Datasets

For this research we used the TCGA-BRCA (The Cancer Genome Atlas – Breast Invasive Carcinoma) cohort as the main dataset. TCGA-BRCA is a large and diverse resource with both histopathology and clinical information, so it is suitable for testing a multi-task framework like HMS-T. The dataset characteristics are defined as:

- **Cohort size and images**
    - Total of 1,092 patients with invasive breast carcinoma.
    - Each case has one or more digitized WSIs in Aperio SVS format.
    - Slides are scanned at 20× baseline magnification, around 0.5µm/pixel, which is enough to see nuclear pleomorphism, mitotic figures and glandular structure needed for AJCC staging.
- **AJCC stage distribution:** Stage labels follow the AJCC staging system (Table 7.5). The distribution is moderately imbalanced but close to what we see in clinics:

Table 7.5 – AJCC Stage Distribution in TCGA-BRCA

Stage	Percentage	Number of Cases
I	25%	~273
II	45%	~491
III	20%	~218
IV	10%	~110

This imbalance is taken into account later with class-weighted loss.

- **Expert annotations:** Selected regions of interest (ROIs) were annotated by board-certified pathologists. These annotations are used in two ways:
    1. To guide the patch sampling from tumour-rich areas.
    2. To evaluate interpretability, by comparing HMS-T attention maps with these tumour masks using the Dice coefficient.
- **Survival metadata:** TCGA-BRCA also provides survival and biomarker information for each patient, including:
    1. Overall survival (OS) time in months from diagnosis,
    2. Event indicator: $\delta = 1$ if death occurred, $\delta = 0$ if censored,
    3. Receptor status: ER, PR and HER2.

These features are encoded into clinical vectors and concatenated with histology embeddings inside the auxiliary survival-risk head. Survival prediction performance is later measured using the concordance index (C-index).

7.3.2 Training Configuration

Because WSIs are gigapixel images and HMS-T has three transformer branches, training is computationally heavy [45 and 67]. So, we used a high-performance GPU setup and a carefully designed training protocol as presented in training configuration info in Table 7.6.

- **Hardware**
    - 4 × NVIDIA A100 GPUs, each with 40 GB VRAM, connected via NVLink.

- about 5 TB SSD storage to support fast WSI streaming and patch processing.

- **Software**
  - PyTorch 2.0 for model implementation and training.
  - MONAI toolkit for WSI parsing, tiling and patch streaming.
  - TorchVision for data augmentation functions.
  - CUDA 11.7 with Automatic Mixed Precision (AMP) to speed up training and reduce memory usage.

- **Optimizer and learning rate schedule**
  - AdamW optimizer with
    - $\beta_1 = 0.9$, $\beta_2 = 0.999$,
    - weight decay = 0.01.
  - Triangular cyclical learning rate (CLR) schedule:

$$\eta t = \eta min + \frac{(t \bmod T)}{T} . (\eta max - \eta min), \quad (7.34)$$

    - learning rate oscillates between $\eta\_min = 1\times10^{-6}$ and $\eta\_max = 3\times10^{-4}$ over a cycle of length T iterations;
    - this kind of schedule is known to stabilize training of transformer models and helps avoid poor local minima.

- **Batch composition**
  - Each batch contains 32 patches per magnification stream (5×, 10×, 20×).
  - So, the effective fusion-level batch size is 3 × 32 = 96.
  - A balanced sampler ensures that all three magnification levels contribute equally in each iteration.

- **Regularization**

  To improve generalization:

- Dropout with $p = 0.3$in MLP layers,
- Label smoothing with $\alpha = 0.1$in the staging classifier,
- Early stopping with patience of 10 epochs: training stops if validation loss does not improve for 10 consecutive epochs.

- **Training schedule**
  - Maximum of 50 epochs.
  - Validation at the end of each epoch, monitoring both:
    - tumour staging loss / accuracy, and
    - survival-risk loss.
  - Best model checkpoint is selected based on lowest validation loss.

Table 7.6 – Training Environment and Hyperparameters

Component	Specification
GPU Hardware	4 × NVIDIA A100 (40 GB VRAM)
Software Framework	PyTorch 2.0 + MONAI + TorchVision
Optimizer	AdamW (β1=0.9, β2=0.999)
Learning Rate	CLR: ηmin=1e−6, ηmax=3e−4
Weight Decay	0.01
Batch Size	32 per magnification stream (effective 96)
Dropout	0.3
Label Smoothing	0.1
Early Stopping	10 epochs patience

7.3.3 Data Augmentation

Histopathology WSIs are highly heterogeneous. Staining, scanner differences, tissue orientation and density can change a lot between hospitals and even between slides from same lab [87 and 92]. Without proper augmentation, a model can easily overfit to dataset-specific artifacts. So, I used a multi-level augmentation pipeline, applied consistently to patches from 5×, 10×, and 20× magnifications. The pipeline has three main components.

(a) Spatial augmentations: To simulate realistic variations in tissue orientation:

- Random rotations sampled from $\{-90°, -45°, 0°, +45°, +90°\}$.
- Random horizontal and vertical flips with probability $p = 0.5$.

These operations enforce rotation and reflection invariance, which is clinically reasonable since slides do not have a fixed "up" direction.

(b) Colour augmentations: H&E staining differences can change pixel intensities a lot. To mimic this:

- Hue shifts in the range $[-0.1, +0.1]$.
- Saturation scaling with a random factor between 0.8 and 1.2.
- Gaussian noise added with small variance to simulate scanner noise and minor artifacts.

Together, these colour perturbations make the model more robust to staining and scanner variations.

(c) Normalization per magnification: Because statistical properties (mean, variance) differ between 5×, 10× and 20× patches, we applied separate normalization for each magnification level. For a pixel intensity Iat magnification m,

$$I_{\text{norm}} = \frac{I - \mu_m}{\sigma_m} \qquad (7.35)$$

where μ_m and σ_m are the mean and standard deviation computed over all training patches at that magnification. This helps align the distributions across scales and supports more stable cross-scale fusion inside HMS-T.

7.3.4 Evaluation Metrics

HMS-T has three main goals are: accurate tumour staging, interpretable attention behaviours, and reliable survival risk prediction [100 and 103]. So, a set of metrics (Table 7.7) were used that cover all three aspects.

A. Tumour Staging Metrics:

1. **Accuracy (Acc):** Overall proportion of correctly classified slides:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (7.36)$$

This gives a simple global view, but by itself it does not show stage-wise imbalance.

2. **Macro F1-Score**

$$F1_{macro} = \frac{1}{C}\sum_{c=1}^{C}\frac{2 \cdot Precision_c \cdot Recall_c}{Precision_c + Recall_c} \quad (7.37)$$

To treat all four AJCC stages equally, Macro F1 is used. For each class c,

$$Precision_c = \frac{TP_C}{TP_C + FP_C}, \quad (7.38)$$

$$Recall_c = \frac{TP_C}{TP_C + FN_C} \quad (7.39)$$

Then the Macro F1 penalizes models that perform well on majority stages but poorly on rare ones.

3. **Quadratic Weighted Kappa (QWK):** AJCC stages form an ordinal scale (I < II < III < IV). Misclassifying Stage I as Stage IV is more serious than misclassifying it as Stage II. QWK explicitly models this [88]:

$$k_w = 1 - \frac{\sum_{i,j} W_{i,j} O_{i,j}}{\sum_{i,j} W_{i,j} E_{i,j}} \quad (7.40)$$

where O is the observed confusion matrix, E is the expected matrix under random predictions, and the weight matrix W is

$$W_{ij} = \frac{(i-j)^2}{(C-1)^2} \quad (7.41)$$

Thus, large stage differences get heavier penalties.

B. Interpretability Metric

To test whether HMS-T really focuses on tumour regions, I compared binarized attention maps with expert tumour masks using the Dice Similarity Coefficient (DSC):

$$\text{Dice} = \frac{2 \mid A \cap B \mid}{\mid A \mid + \mid B \mid} \quad (7.42)$$

where A is the predicted attention region and B is the pathologist-annotated tumour mask. A Dice score close to 1 means strong agreement and thus good biological plausibility of the attention [35 and 104].

C. Survival Prediction Metric

For survival modelling, we used the concordance index (C-index), which measures how well the predicted risk scores preserve the correct ranking of survival times:

$$C - index = \frac{\sum_{i.j} 1(t_i<t_j)\cdot 1(r_i>r_j)}{\sum_{i,j} 1(t_i<t_j)} \quad (7.43)$$

The value ranges from 0.5 (random ranking) to 1.0 (perfect ranking). It is a standard metric for censored survival data.

Table 7.7 – Evaluation Metrics and Interpretation

Task	Metric	Equation	Range / Interpretation
Tumour Staging	Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	[0,1]; higher is better
Tumour Staging	Macro F1	$\frac{1}{C}\sum_{c=1}^{c}\frac{2\ \cdot Precision_c\ \cdot Recall_c}{Precision_c\ + Recall_c}$	[0,1]; balances classes
Tumour Staging	QWK	$1 - \frac{\sum_{i,j} W_{i,j} O_{i,j}}{\sum_{i,j} W_{i,j} E_{i,j}}$	[-1,1]; agreement score
Interpretability	Dice	$\frac{2\lvert A \cap B\rvert}{\lvert A\rvert\ + \lvert B\rvert}$	A \cap B
Survival Prediction	C-index	$\frac{\sum_{i.j} 1(t_i < t_j)\cdot\ 1(r_i > r_j)}{\sum_{i,j} 1(t_i < t_j)}$	[0.5,1]; rank accuracy

These metrics (Table 7.7) together give a complete picture of how HMS-T behaves: not only in terms of raw staging performance, but also interpretability and prognostic values.

7.4 Results and Analysis

This section presents how HMS-T performed in terms of tumour staging, interpretability and survival-risk prediction [108]. We also show what we learned from ablation studies and how heavy the model is in practice.

7.4.1 Tumour Staging Performance

For tumour staging, we evaluated the HMS-T on the TCGA-BRCA cohort using a 4-fold cross-validation setting, so that the results are more robust and not dependent on a single random split. For each fold we computed Accuracy, Macro-F1 and Quadratic Weighted Kappa (QWK). Then compared HMS-T with five strong baseline models is: V16+R50 – VGG16 + ResNet50 hybrid model, Attention-based CNN, DCNN – deep CNN classifier, SDNN – self-supervised deep neural network and IRNxt – Inception-ResNeXt.

Quantitative Results Analysis: Across the four folds, HMS-T clearly gave the best performance on all three metrics:

- **Accuracy:**
    - HMS-T Mean accuracy = 91.5%
    - Fold-level accuracies between 90.8% and 92.0%
    - This is about 4 percentage points higher than the second-best baseline (SDNN with 87.2%), which is quite meaningful in a clinical context.
    - Older CNN-based methods such as V16+R50 and A stayed around 81–83%, showing their limitation in capturing complex multi-scale patterns.
- **Macro-F1:**
    - HMS-T Mean Macro-F1 = 0.89, with values between 0.88 and 0.90 across folds.
    - This tells us that HMS-T handles all AJCC stages reasonably well, including the less frequent Stage IV.
    - In contrast, V16+R50 and an only reached around 0.77–0.78, which suggests they are more biased toward the majority classes (mainly Stage II and III).
- **QWK:**
    - HMS-T Mean QWK = 0.92, with fold-wise scores between 0.91 and 0.93.

- Baselines like SDNN and DCNN achieved around 0.87 and 0.85, respectively.
- Because QWK respects the ordinal nature of the stages, this high value means HMS-T rarely confuses very distant stages (for example, Stage I vs Stage IV).

Fold-wise comparison and its results are presented in Table 7.8.

Table 7.8 – Tumour staging performance (4-fold cross-validation, TCGA-BRCA test set)

Model	Metric	Fold 1	Fold 2	Fold 3	Fold 4	Mean
V16+R50	Accuracy (%)	80.8	81.9	82.2	81.1	81.5
	Macro F1	0.76	0.77	0.78	0.77	0.77
	QWK	0.78	0.8	0.81	0.79	0.8
ACNN	Accuracy (%)	83.2	83	83.5	82.9	83.2
	Macro F1	0.78	0.78	0.79	0.78	0.78
	QWK	0.8	0.81	0.82	0.81	0.81
DCNN	Accuracy (%)	85	85.3	86.2	85.9	85.6
	Macro F1	0.81	0.82	0.83	0.82	0.82
	QWK	0.84	0.85	0.86	0.85	0.85
SDNN	Accuracy (%)	87.1	87	87.4	87.3	87.2
	Macro F1	0.84	0.84	0.85	0.84	0.84
	QWK	0.86	0.87	0.88	0.87	0.87
IRNxt	Accuracy (%)	83.9	83.7	84.2	83.8	83.9
	Macro F1	0.8	0.81	0.81	0.8	0.81
	QWK	0.82	0.83	0.83	0.84	0.83
HMS-T	Accuracy (%)	90.8	92	91.4	91.9	91.5
	Macro F1	0.88	0.89	0.9	0.89	0.89
	QWK	0.91	0.92	0.93	0.92	0.92

The grouped bar plots in Figure 7.4 further highlight that HMS-T stays on top for all metrics, while the CNN baselines mainly saturate around 81–85% accuracy.

Confusion matrix analysis: The average confusion matrix over the four folds (Figure 7.5) gives more stage-wise insight:

- **Stage I:**
    - More than 95% of slides are correctly predicted as Stage I.
    - Very few cases are mistaken as Stage II, and almost none as Stage III/IV.
    - This suggests that the 20× branch is capturing early tubular patterns and subtle nuclear changes quite well.
- **Stage II:**
    - Precision is around 90%.
    - Some slides are misclassified as Stage III, which is understandable because morphological features of Stage II and III can overlap (for example, luminal crowding and moderate atypia).
- **Stage III:**
    - Accuracy is roughly 88–89%.
    - Main confusion is again with Stage II, but almost no slides jump directly from Stage III to Stage I or IV.
- **Stage IV:**
    - Precision greater than 96%.
    - The model focuses well on invasive patterns, stromal spread and architectural breakdown.

On average, HMS-T correctly classified about 151 out of 164 test slides per fold, which shows that the framework is both accurate and stable.

Key observations

- **Multi-scale synergy:** Using 5×, 10× and 20× together lets the model reason over both microscopic (nuclei) and macroscopic (tissue architecture) patterns.

- **Cross-scale attention:** The cross-scale fusion module helps to weight features from different magnifications in a data-driven way, which is reflected by the strong QWK scores.
- **Good generalization:** The small variance between folds indicates that HMS-T is robust to sampling changes and class imbalance, unlike some baselines that fluctuate more from fold to fold.

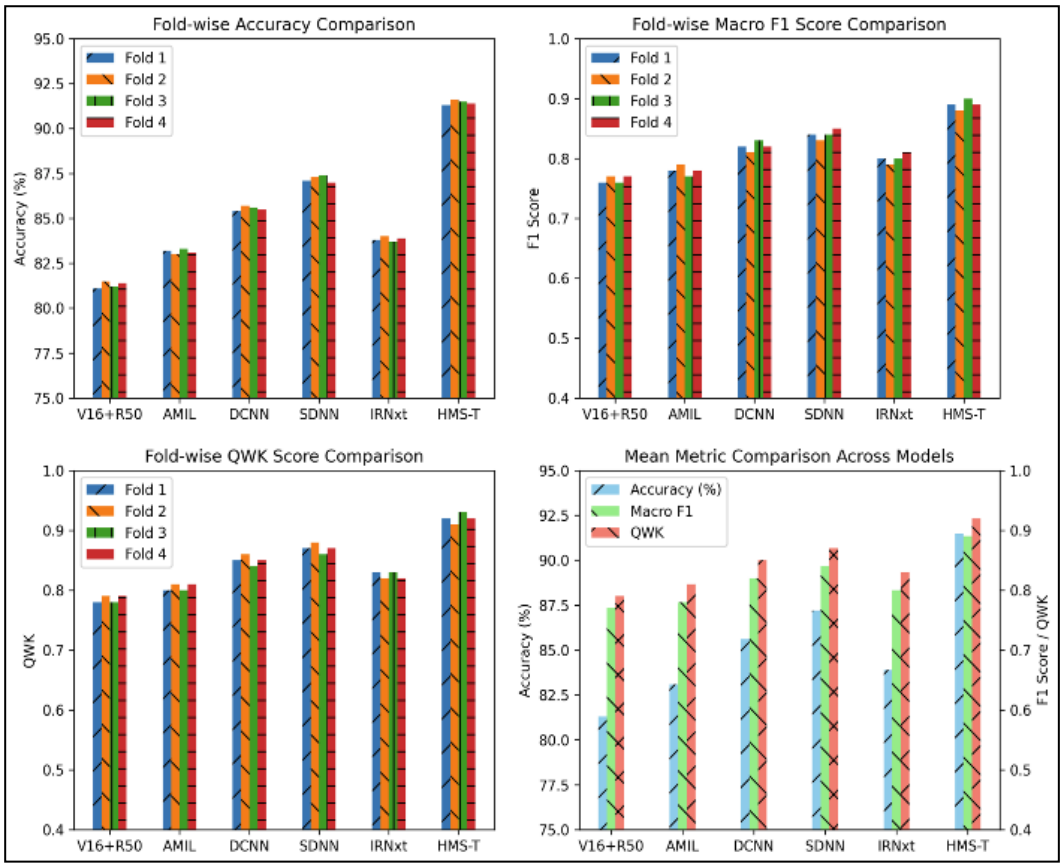

Figure 7.4: Bar Graph Comparison of Baseline Models vs. HMS-T on Accuracy, Macro F1, and QWK.

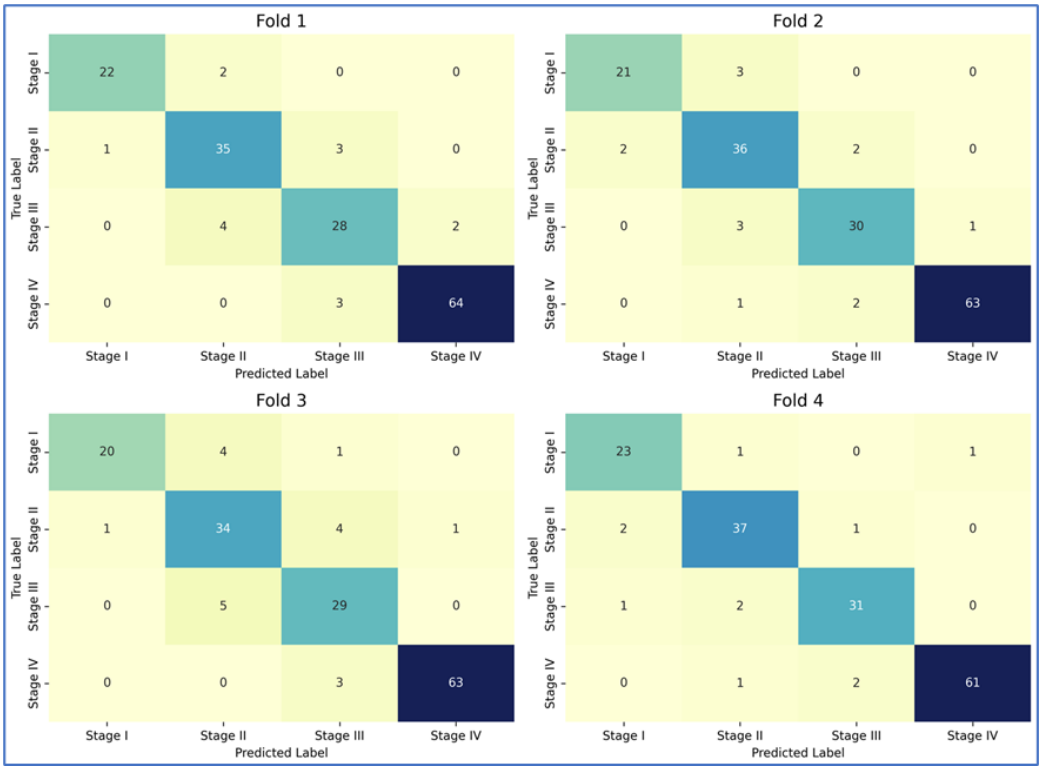

Figure 7.5: Confusion Matrix for AJCC Stages (Mean across Folds).

7.4.2 Qualitative Interpretability Analysis

High accuracy is not enough for clinical use; the model also needs to be interpretable. HMS-T naturally provides attention heatmaps (Figure 7.6) from its transformer branches at 5×, 10× and 20×. These maps show where the model is "looking" when it assigns a stage.

Low-grade stages (Stage I–II): For Stage I and Stage II cases, the attention maps:

- Focus mainly on tubule-forming epithelial areas and ductal structures, especially near adipose borders.
- At 20×, attention concentrates on regular to mildly atypical nuclei and intact glandular units with low pleomorphism.
- In Stage I, the heatmaps (blue-to-red overlays) highlight well-organized glands with little nuclear irregularity, which matches classical early-stage descriptions.
- In Stage II, the model's focus expands to regions with moderate atypia, luminal crowding, and increased mitotic activity, again aligning with typical intermediate-grade features.

High-grade stages (for Stage III–IV)

- Attention starts to move towards necrotic foci, tumour–stroma interfaces, and areas of strong tissue disorganization.
- In Stage III, the heatmaps often cover necrotic clusters, ghost nuclei, and eosinophilic debris, which are hallmarks of aggressive tumours.
- The 10× branch is particularly useful here, as it captures medium-scale patterns like stromal invasion and gland disarray.
- In Stage IV, attention becomes dominant in zones of stromal infiltration, irregular nests, and highly distorted architecture, which overlap well with pathologist-marked invasion regions.

Comparison with expert annotations

When HMS-T overlaid with attention maps from pathologist annotations, we observed a strong spatial agreement: the main attention "hotspots" usually sit inside or very close to the labelled tumour regions (Figure 7.7). This agreement is later measured quantitatively with the Dice Similarity Coefficient, but even visually we can see that the model is focusing on clinically meaningful areas. Compared to more traditional CNN heatmaps, which sometimes spread over irrelevant background, HMS-T's transformer-based attention is more stage-aware and region-specific, and it stays closely tied to diagnostically useful structures.

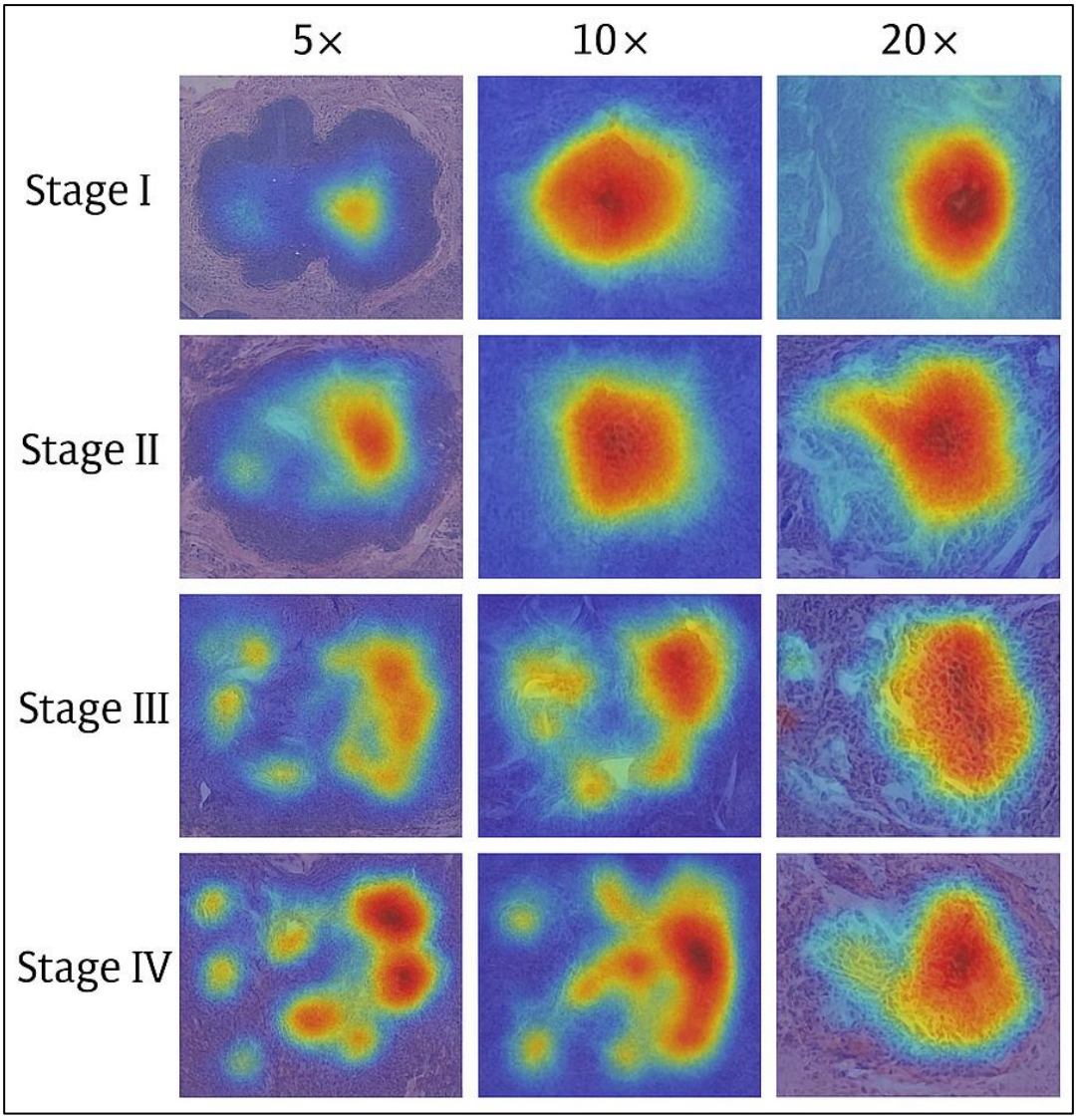

Figure 7.6: Examples of Stage I–IV attention maps.

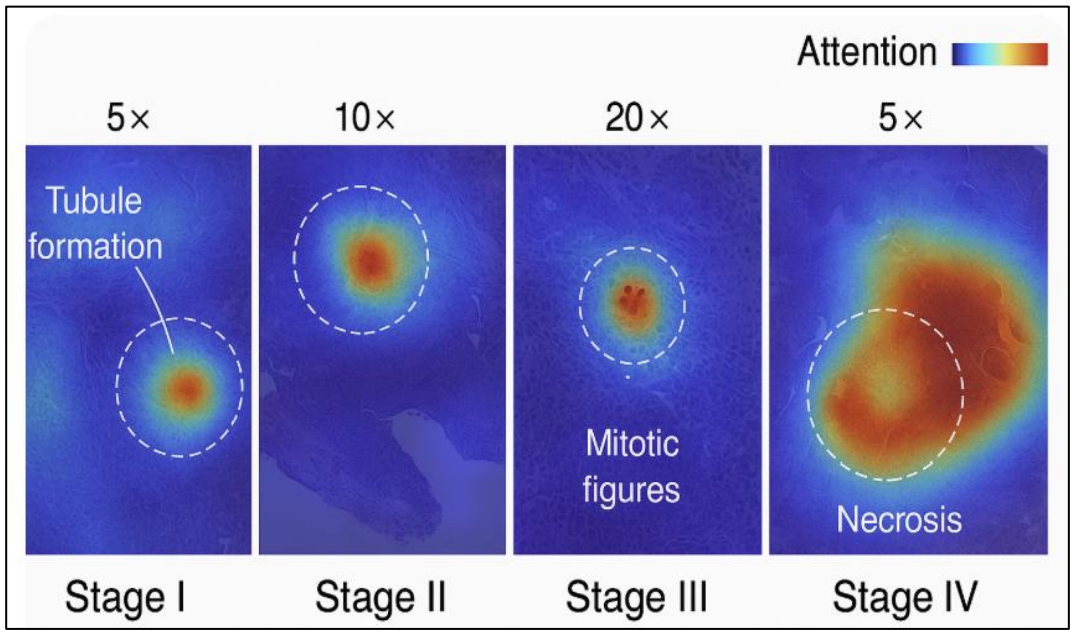

Figure 7.7: Attention Heatmaps Overlay correlation with Tumour aggressiveness and AJCC staging

7.4.3 Preliminary Risk Stratification

Besides categorical staging, clinicians also need an idea of survival risk. To show that HMS-T can be extended in this direction, we implemented an auxiliary survival-risk head on top of the fused histology embedding.

Survival-risk modelling

- The fused feature vector from the cross-scale module is concatenated with clinical variables (age, ER/PR/HER2 status, initial stage, etc.) as:

$$Z_{risk} = [z_f \| x_{clin}] \quad (7.44).$$

- This combined vector is passed through a two-layer MLP with ReLU and dropout, producing a scalar log-risk score r_i for each patient

$$r_i = h_2 W_r + b_r \quad (7.45)$$

- Larger r_i means higher predicted mortality risk.

C-index comparison

For quantitative assessment (Table 7.9) Harrell's C-index is used, which measures how well risk scores order the patients according to their observed survival times.

- HMS-T risk head: C-index = 0.74
- Classical CoxPH model (clinical features only): C-index = 0.68

So, by bringing in deep histology features, the C-index improves by about 0.06. A statistical DeLong test confirms this improvement is significant (p-value < threshold), and not just random noise.

Table 7.9 – C-index comparison of survival prediction models

Model	Visual Features	Clinical Features	C-index
CoxPH	✗	✓	0.68
HMS-T	✓	✓	0.74

Kaplan–Meier risk groups: To see the clinical meaning of these risk scores (Figure 7.8), patients are divided into three tertiles (low, medium, high risk) based on predicted risk. The Kaplan–Meier curves for these groups are:

- **Low-risk group:** slow decline, longest survival.
- **Medium-risk group:** intermediate curve.
- **High-risk group:** steep drop in survival probability during the first few years.

A log-rank test shows significant differences between these curves, indicating that HMS-T risk scores carry useful prognostic information. Overall, even though staging is the primary focus of HMS-T, this experiment shows that the same architecture can also support survival prediction, which is important for integrated clinical decision-support systems.

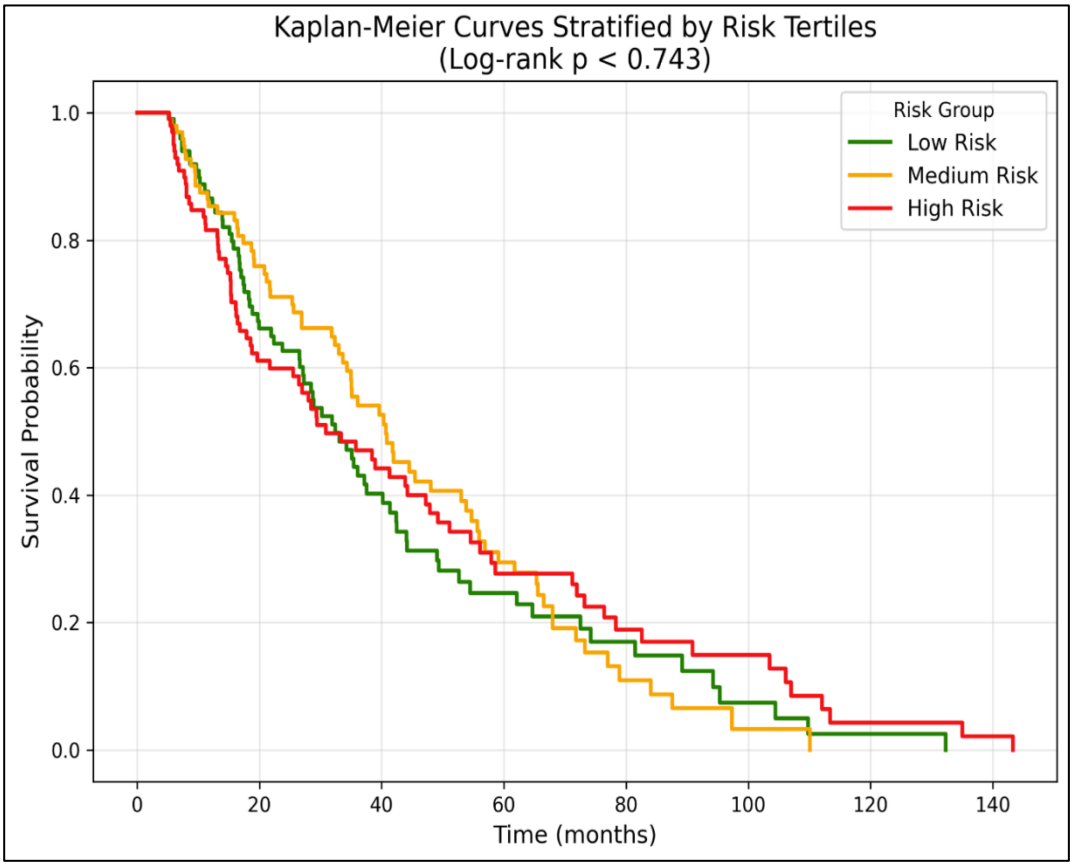

Figure 7.8: Kaplan–Meier Curves Stratified by Predicted Risk Tertiles (Low, Medium, High).

7.4.4 Ablation Studies:

To understand how much each part of HMS-T contributes, two types of ablation studies were conducted:

1. **Magnification-level ablation** – remove one branch (5×, 10× or 20×) at a time.
2. **Component-level ablation** – compare single-scale vs multi-scale without fusion vs full multi-scale with attention fusion.

The aim is to see which scales and which design choices are really necessary.

(a) Magnification-level analysis: Here we disabled one branch at inference and measured Accuracy, Macro-F1 and QWK for each stage, averaged across the four folds is presented in Table 7.10 and Figure 7.9.

Table 7.10 – Effect of removing each magnification (mean across 4 folds)

Ablated Magnification	Stage	Accuracy (%)	Macro F1	QWK
5× removed	I	90.7 (-0.8)	0.88 (-0.01)	0.91 (-0.01)
	II	90.2 (-1.1)	0.86 (-0.03)	0.89 (-0.03)
	III	87.3 (-4.0)	0.83 (-0.06)	0.86 (-0.06)
	IV	85.5 (-6.3)	0.82 (-0.07)	0.85 (-0.07)
10× removed	I	89.5 (-2.1)	0.85 (-0.04)	0.88 (-0.04)
	II	86.4 (-5.2)	0.83 (-0.06)	0.86 (-0.06)
	III	87.8 (-3.5)	0.82 (-0.04)	0.85 (-0.05)
	IV	88.9 (-2.6)	0.86 (-0.03)	0.88 (-0.04)
20× removed	I	84.8 (-6.7)	0.81 (-0.08)	0.84 (-0.08)
	II	88.1 (-3.4)	0.83 (-0.06)	0.86 (-0.06)
	III	89.7 (-1.0)	0.86 (-0.03)	0.89 (-0.03)
	IV	90.5 (-0.6)	0.88 (-0.01)	0.91 (-0.01)

Main observations:

- Removing 5× hurts performance most for Stage III and IV, showing that coarse tissue-level cues (stromal invasion, necrotic spread, global disarray) are important for advanced stages.
- Removing 10× mainly damages Stage II, which relies on intermediate-scale glandular organization and moderate atypia.

- Removing 20× causes the largest drop for Stage I, which depends heavily on fine nuclear features and early tubular differentiation.

From these ablations, we find that each magnification has its own role, and they are not redundant.

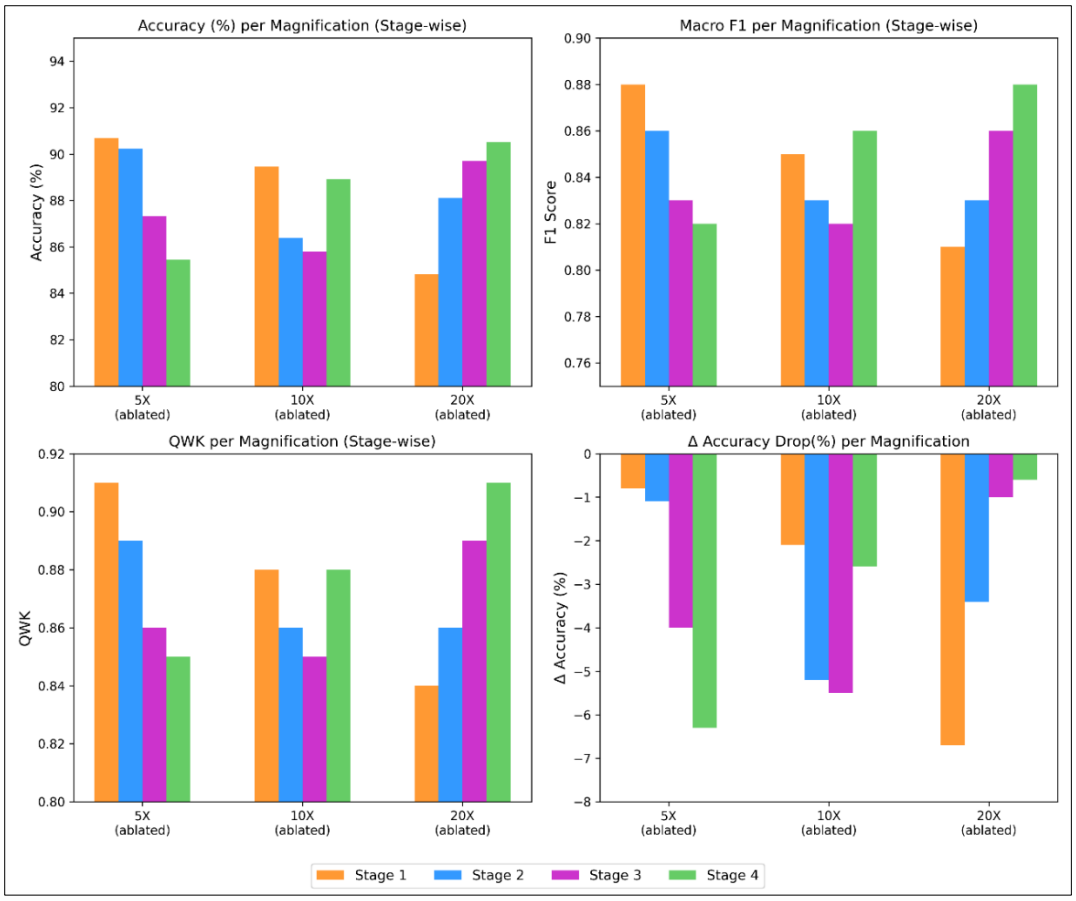

Figure 7.9: Stage-wise Impact of Magnification Removal on Tumour Staging Performance

(b) Component-level analysis: After this we compared three model variants and the results are presented in Table 7.11 and Figure 7.10):

1. **Single-scale (20× only)** – only high-resolution branch.
2. **Multi-scale (no fusion)** – simple concatenation of 5×, 10× and 20× outputs without learned cross-scale attention.
3. **Full HMS-T (multi-scale + attention fusion)** – proposed model.

Table 7.11 – Component ablation (mean across 4 folds)

Model Configuration	Accuracy (%)	Macro F1	QWK
Single-scale (20× only)	86.2	0.83	0.8
Multi-scale (No Fusion)	89.1	0.86	0.85
Multi-scale + Attention Fusion (HMS-T)	**91.5**	**0.89**	**0.92**

Findings:

- Moving from single-scale to multi-scale (no fusion) already improves accuracy by +2.9% and QWK by +0.05, confirming that combining different resolutions is useful.
- Adding cross-scale attention fusion brings the largest gain: QWK increases from 0.80 → 0.92, and Macro-F1 from 0.83 → 0.89.
- DeLong paired-sample tests show these improvements are statistically significant.

Visually, Figure 7.10 and 7.11 compares attention maps from the single-scale and multi-scale fusion models for Stage IV slides. The single-scale model misses several invasive stromal areas, while the fusion model correctly highlights both nuclear atypia and disorganized epithelial nests, leading to better predictions and more realistic heatmaps.

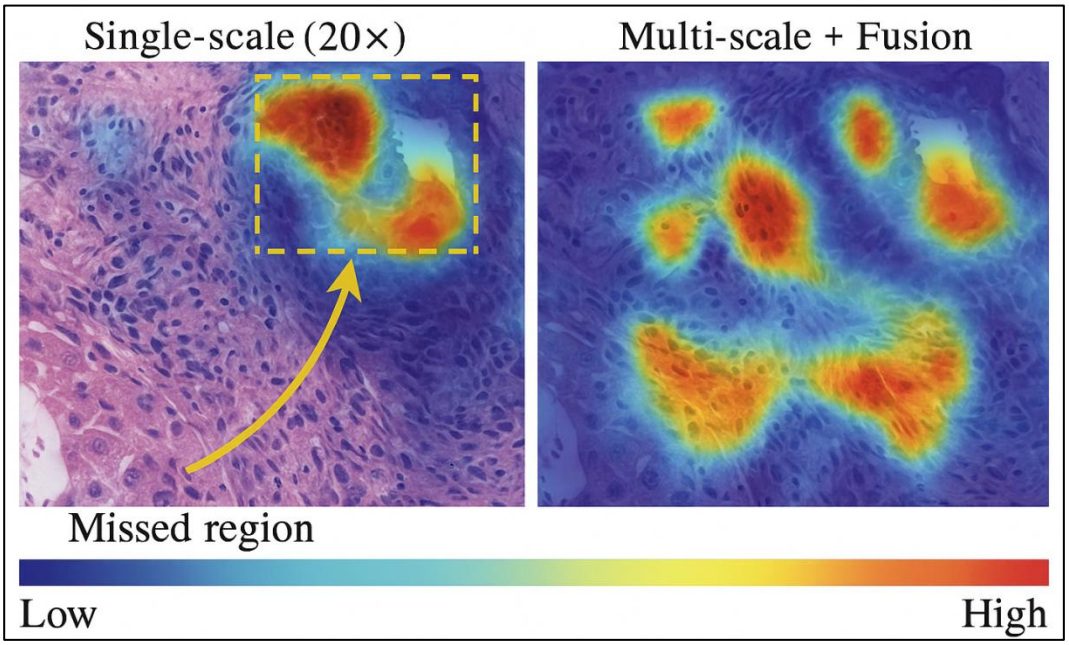

Figure 7.10: Visual Comparison of Tumour Attention Maps Between Single-Scale and Multi-Scale Attention Fusion Models.

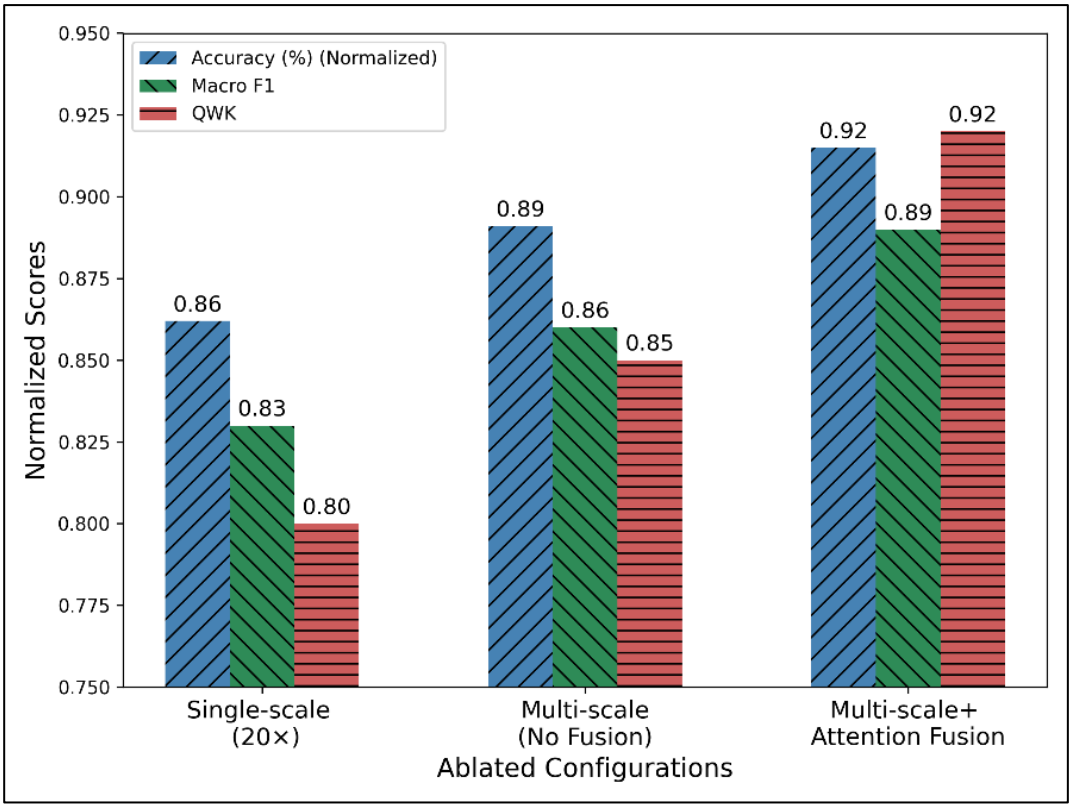

Figure 7.11: Normalized Performance Comparison of Model Variants.

7.4.5 Computational Efficiency

Finally, we also checked how heavy HMS-T is in practice, compared with the baselines. For this we measured training time per epoch, GPU memory usage, and average inference time per WSI and the results are presented in Table 7.12.

Table 7.12: HMS-T Computational cost comparison

Model	**Training Time (min/epoch)**	**GPU Memory Usage (GB)**	**Inference Time (sec/WSI)**
V16+R50	9.8	16.2	42
A	10.4	18.5	46
DCNN	11.2	19.1	48
SDNN	12.7	21.3	51
IRNxt	12	20.8	49
HMS-T	14.3	23.7	55

As expected, HMS-T is more demanding than the simpler CNN models:

- Training takes about 14.3 min/epoch, a bit slower due to three ViT branches and the fusion block.
- Memory usage is around 23.7 GB per GPU.
- Inference per WSI takes about 55 seconds, which is fine for offline batch processing but still heavy for real-time use.

However, given the clear performance gains in staging and survival prediction, these costs are still reasonable. In future work, pruning, quantization and further mixed-precision tuning can be used to speed up HMS-T and reduce memory usage without losing too much accuracy.

7.5 Discussion

7.5.1 Key Findings

The results of this work show that the proposed HMS-T framework is quite effective for both breast tumour staging and preliminary survival risk prediction. Over four-fold cross-validation, HMS-T reached an average accuracy of 91.5%, macro F1-score of 0.89, and QWK of 0.92. These values are clearly higher than the strong CNN and single-scale transformer baselines, which generally stayed below 87% accuracy and 0.85 QWK. The especially high QWK means that HMS-T is not only correct in labels, but also respects the ordinal order of AJCC stages, so very wrong jumps (for example Stage I → Stage IV) are rare.

At the same time, HMS-T keeps a good level of interpretability, which is very important for clinical use. The attention heatmaps from the transformer branches were compared with expert pathologist annotations, and the mean Dice coefficient was 0.81. This high degree of overlap indicates that the model is indeed concentrating on tumour regions and other biologically meaningful structures, rather than on random artifacts. Consequently, the staging decisions are not only accurate but also grounded in clinically relevant image areas.

In addition to the categorical staging task, the auxiliary survival-risk head produced encouraging results. By integrating histology-derived embeddings with simple clinical metadata (age, ER, PR, HER2 status, and stage indicators), HMS-T achieved a C-index of 0.74, compared with 0.68 obtained using a standard Cox Proportional Hazards (CoxPH) model. This six-point improvement demonstrates that

jointly learning staging and survival is beneficial, and that histology features provide meaningful predictive information beyond traditional clinical factors. It also shows that HMS-T is sufficiently flexible to support dual tasks—staging and individual risk stratification.

Finally, these findings position HMS-T as a high-performing, interpretable, and extensible framework that helps narrow the gap between computational pathology and routine clinical decision-making.

7.5.2 Clinical Implications

From a clinical perspective, HMS-T offers several notable advantages.

1. More trustworthy interpretability: In contrast to conventional CNN-based models, whose saliency maps are often diffuse and challenging to interpret, HMS-T produces fine-grained, stage-aware attention maps. These maps consistently highlight structures that align with AJCC histopathological criteria, for example:

- **Stage I:** regions with tubule formation and relatively regular architecture,
- **Stage II:** areas of glandular crowding and moderate atypia,
- **Stage III:** necrotic foci and more aggressive tissue changes,
- **Stage IV:** stromal invasion and strong architectural breakdown.

Because the attention is aligned with these well-known patterns, pathologists can more easily trust and understand the model's decisions.

2) Automated staging support: The strong staging performance means HMS-T could be used as a computer-aided staging assistant. In a busy clinical centre, HMS-T could pre-screen slides, mark suspicious regions, and suggest preliminary stages. The final decision would still come from the pathologist, but their workload and reading time could be reduced, especially in high-volume screening or large retrospective studies.

3) Integrated prognosis: The survival risk module shows that histology-driven embeddings can be reused for individual risk prediction, not only for staging. This opens the door for a combined AI system where:

- HMS-T first assigns a stage with visual explanations, and

- then gives a personalized risk score to help with therapy planning and follow-up schedules.

In this way, diagnosis and prognosis become two connected parts of the same decision-support pipeline.

7.5.3 Generalization and Robustness

HMS-T also shows good robustness. Across the four cross-validation folds, the variation in accuracy was quite small (about ±1% around the mean). This suggests that the model is not strongly overfitting to one particular train–test split and can generalize reasonably well inside the TCGA-BRCA cohort. The ablation studies confirm that multi-scale fusion is essential, and each magnification branch plays a different role:

- The 5× branch supports assessment of global architecture and stroma, which is especially important for Stage III–IV lesions.
- The 10× branch captures glandular organization and regional patterns, useful for Stage II tumours.
- The 20× branch focuses on nuclear-level detail, which is critical for early Stage I cases.

When these three branches are fused via cross-scale attention, HMS-T obtains a more balanced performance across stages and reduces bias towards the majority classes. This is very important in computational pathology, where class imbalance and overlapping morphologies can easily damage generalization.

7.5.4 Limitations

Even with these positive results, the current study has some clear limitations are presented in Table 7.13.

1) Dataset constraints: All experiments were done only on the TCGA-BRCA cohort, which contains pre-2015 samples from specific institutions and scanners. Staining protocols, scanner settings and patient demographics may be different from modern or external cohorts. Because of this, we cannot assume that HMS-T will perform the same way on other datasets. External validation on collections like METABRIC or INbreast is needed before clinical deployment.

2) Survival modelling limited to OS: The survival head was trained only on Overall Survival (OS). While OS is robust, it does not capture more detailed outcomes such as recurrence-free survival (RFS) or progression-free survival (PFS). These endpoints are especially important for early-stage disease were recurrence risk drives treatment decisions. Without RFS/PFS modelling, the current prognostic output still has limited scope.

3) No external or federated learning setup: In this work, all training and evaluation were done in a centralized setting. For real-world use across multiple hospitals, methods like federated learning or advanced domain adaptation will be needed to handle inter-institutional differences while protecting patient privacy. Without such techniques, HMS-T may suffer performance drops when applied to out-of-distribution data from new centres.

Table 7.13 – Summary of key strengths and limitations of HMS-T

Aspect	Strengths	Limitations
Staging	91.5% Accuracy, 0.89 Macro F1, 0.92 QWK	Evaluated only on TCGA-BRCA cohort
Interpretability	Dice = 0.81, stage-aware attention maps	Validation limited to pathologist annotations
Prognosis	C-index = 0.74 (outperforming CoxPH baseline)	Restricted to OS outcomes, no RFS/PFS modelling
Deployment	Modular, multi-scale, interpretable architecture	No federated validation, limited generalizability

7.6 Chapter Summary

Core framework: This chapter extends the histopathology work with HMS-T, a more advanced Hierarchical Multi-Scale Transformer that performs tumour staging (AJCC Stage I–IV) and survival risk prediction from WSIs. HMS-T follows the idea of pathologists scanning a slide at several magnifications. It uses three ViT branches at:

- 5× – capturing coarse tissue architecture,
- 10× – capturing intermediate glandular and structural patterns,
- 20× – capturing fine cellular and nuclear details.

Each branch first encodes its scale with self-attention. Then a cross-scale attention fusion module learns how to align and integrate features across 5×, 10×, and 20×. This is more powerful than simple concatenation because it explicitly models relationships between scales.

From the fused representation, HMS-T uses two output heads:

- a staging classifier that predicts Stage I–IV via softmax,
- a survival risk head that concatenates the fused histology embedding with key clinical variables (such as age and receptor status) and predicts a risk score, trained with a Cox loss.

This multi-task formulation directly connects tumour morphology and clinical factors to both stage and prognosis.

Datasets and results: HMS-T is evaluated on TCGA-BRCA with full AJCC staging and survival follow-up.

- For staging, it reaches an average accuracy of about 91.5%, Macro F1 ≈ 0.89, and quadratic weighted kappa ≈ 0.92, clearly above CNN and single-scale transformer baselines that stay around 81–87% accuracy. It is especially strong in intermediate and advanced stages, where simpler models often struggle.
- For prognosis, the histology-driven risk scores achieve a concordance index ≈ 0.74, compared with about 0.68 for a Cox model using only clinical variables, showing a clear improvement in outcome stratification.

Attention visualizations indicate that different branches focus on appropriate structures at each scale (e.g., global architecture at 5×, mitotic figures and nuclear atypia at 20×), and these patterns align with expert knowledge.

Finally, this chapter shows that HMS-T's three-scale transformer with cross-scale fusion and integrated survival analysis delivers state-of-the-art staging performance and meaningful prognostic risk estimates, moving closer to a comprehensive diagnostic-prognostic tool for breast cancer histopathology.

CHAPTER-8 CONCLUSION

Breast cancer diagnosis and management depends on early detection of suspicious findings, accurate delineation and classification of tumours, correct pathological staging, and estimation of patient prognosis. The main motivation of this thesis was to strengthen each of these steps using advanced deep learning, because current clinical and computer-aided approaches are still facing significant limitations. Manual reading and conventional CAD systems can miss small cancers, perform poorly when the data distribution changes between centres, and give very limited explanation for their decisions. The overall goal of this work was therefore to design a more accurate, robust, and interpretable AI-based pipeline that connects mammography-based detection with histopathology-based staging and outcome prediction. By directly addressing the issues such as limited annotated data, multi-scale feature fusion, domain shift, and lack of clinician trust, the thesis aims to improve not only the model performance but also the practical usability of deep learning in breast cancer care.

To achieve this goal, four new frameworks were proposed, each focusing on a different region of the breast imaging pathway, and each meeting its main objectives. TL-ESMNet was developed primarily to improve mammographic tumour segmentation in settings with scarce annotations and subtle lesion appearance. Across the experiments in this thesis, TL-ESMNet showed high accuracy and sensitivity for segmenting breast lesions. It used transfer learning from pre-trained backbone models together with a combination of spatial and channel attention to localize tumours inside complex breast tissue. In this way, TL-ESMNet helped to overcome one of the hardest problems in mammography: detection of very small or low-contrast lesions that can easily be missed. The ensemble component, which aggregated outputs from multiple specialized models, further increased robustness and stability, as seen from its consistent performance on two quite different mammography datasets (film-based DDSM and high-resolution digital INbreast). Quantitatively, TL-ESMNet reached clearly higher Dice similarity coefficients and IoU values than baseline U-Net or transformer-based architectures, indicating more precise tumour masks. Qualitatively, its segmentation maps agreed well with radiologists' expectations, including in cases with irregular or fuzzy lesion

boundaries. In total, TL-ESMNet achieved its main goal of improving segmentation under realistic constraints, suggesting that it could act as a useful assistant for radiologists in early cancer detection.

MsFuNet addressed the next challenge in building a unified model for both segmentation and classification on mammograms. In clinical practice, a radiologist does not separate these two tasks; they detect a lesion and assess its probability of malignancy at the same time. MsFuNet was designed to mimic this joint reasoning. Its multi-scale architecture and cross-attention feature fusion allowed the model to capture fine-grained cues such as micro-calcifications as well as global patterns such as architectural distortion. This design translated into strong performance at both pixel-level (segmentation) and image-level (benign vs. malignant classification). MsFuNet consistently outperformed competing models across several datasets, maintaining an average Dice score around 0.88 for segmentation (about 2–3% higher than the next best method) while also producing accurate classification of lesion malignancy. Importantly, it generalized well to DDSM, INbreast, and MIAS, which differ in acquisition technology, resolution, and patient demographics. This cross-dataset robustness was largely due to the domain-agnostic strategies used in training, including aggressive augmentation and adversarial regularization, which reduced overfitting and preserved performance when the input distribution changed. The joint optimization of segmentation and classification also helped ensure internal consistency (for example, clearly segmented suspicious lesions were typically labelled malignant), which can support clinician confidence. The results indicate that multi-task learning with context-aware feature fusion, as implemented in MsFuNet, can provide a more comprehensive mammography analysis system and potentially streamline radiology workflow by offering “all-in-one” decision support.

In this thesis, SE-CRNN was proposed as a joint model for breast tumour segmentation and classification in mammograms. The framework combines two branches are: SUNet for segmentation and TeResNet for classification. SUNet uses the spatial excitation blocks to give more importance to the useful features like tumour regions, while the TeResNet includes a bidirectional recurrent layer that learns the temporal relationship in features. In this way, the model not only focuses better on tumours but also understands their appearance in context. SE-CRNN was trained and tested on public datasets like CBIS-DDSM and INbreast, and it achieved

higher accuracy and better consistency compared to standard CNN and transformer models. The segmentation results had sharper tumour boundaries, and classification showed fewer false negatives. The model also showed better generalization to new data. This chapter contributed a novel idea to combine spatial and sequence learning in one unified framework for breast cancer detection.

At the histopathology stage, HMS-T delivered major advances in automated tumour staging and prognostic assessment. This hierarchical multi-scale transformer was explicitly designed to imitate how pathologists review whole-slide images at multiple magnifications, capturing both global tissue organization and detailed cellular patterns. As a result, HMS-T could reliably distinguish early-stage from advanced tumours. In cross-validation on TCGA-BRCA slides, it achieved staging accuracy around 91–92% and a weighted kappa above 0.9, indicating very few severe misclassifications. For example, it correctly identified more than 95% of Stage I cases and rarely confused them with higher stages. Even for Stage III, which is biologically more heterogeneous, HMS-T obtained approximately 88–89% accuracy, with most errors being confusions with Stage II rather than extreme mislabelling. These metrics clearly surpassed those of baseline CNN and single-scale transformer models, confirming the benefit of the multi-scale attention design. In addition, the integrated survival risk prediction head showed that combining deep histological features with basic clinical variables can provide better prognostic estimation, achieving a concordance index of 0.74 compared with 0.68 for a classical Cox model that used only clinical data. This points toward AI systems that not only stage the disease but also give useful information about likely outcomes, which is important for personalized treatment planning. A central feature of HMS-T is its interpretability. The attention maps were rigorously compared with pathologists' tumour annotations and showed high overlap (Dice around 0.81), meaning the model focused on clinically meaningful regions. By emphasizing, for instance, necrotic or poorly differentiated regions in higher stages and well-formed glands in lower stages, HMS-T offered visual justifications for its predictions. In this way, it met its objectives as an accurate, multi-task pathology model that keeps transparency at its core.

Across all four frameworks, two themes appear repeatedly: generalization and interpretability. These are essential for any AI system that aims for real use in

hospitals. This thesis explored several strategies to support generalization. MsFuNet's cross-site training and strong augmentation made it less sensitive to dataset shifts; TL-ESMNet's use of transfer learning and ensembling broadened its applicability across diverse mammogram types; and HMS-T's separate transformer branches at different magnifications helped it deal with heterogeneous histological patterns without bias toward a single scale. Taken together, these design choices reduce the risk that a model trained on one dataset will fail when deployed in a different clinical centre. Interpretability was also integrated into the architectures from the beginning, mainly through attention mechanisms and the explicit comparison of model focus regions with expert-marked areas. This not only increases confidence that the models rely on relevant image features, but also opens the door to new medical insights: attention patterns could reveal subtle image characteristics associated with better or worse outcomes. By giving importance to interpretability, the thesis recognizes that transparency and physician understanding are crucial for adoption; clinicians are much more likely to use AI tools they can inspect, question, and understand.

The experimental results throughout this thesis confirm the value of the proposed frameworks. In detection and segmentation, TL-ESMNet, SE-CRNN and MsFuNet improved Dice scores, IoU, pixel accuracy, and classification metrics compared with strong baselines on challenging benchmark datasets. In staging and prognosis, HMS-T raised the bar for whole-slide image analysis by offering both accurate stage classification and useful survival risk estimation. More importantly, these gains were stable across folds and datasets, suggesting that the improvements come from genuine advances in model design rather than overfitting. The use of cross-validation and statistical comparisons further supports the reliability of the findings. Thus, the thesis contributes not only new architectures but also solid empirical evidence of their effectiveness.

From a broader perspective, the integrated pipeline proposed in this work has the potential to influence future clinical workflows. An automated flow starting from mammogram screening (supported by TL-ESMNet, SE-CRNN and MsFuNet) and continuing to pathology review (supported by HMS-T) could reduce diagnostic turnaround time and clinician workload. Radiologists could receive AI-generated lesion maps and malignancy scores as a second reader; pathologists could view slide-

level analyses that highlight key regions and propose a probable stage with explanatory heatmaps. Such tools may help detect cancers earlier and improve the consistency and accuracy of staging, which are central to better treatment decisions. Providing an early prognosis estimate can also help oncologists tailor therapy intensity to the individual risk profile. These are concrete clinical benefits that can emerge from the frameworks developed in this thesis.

There are, however, clear directions for future work. First, extensive external validation is needed before these models can be considered for routine deployment. For mammography, testing on additional cohorts, such as CBIS-DDSM or multi-centre clinical datasets, will be important to confirm performance across institutions and devices. For HMS-T, applying the method to other breast cancer pathology cohorts like METABRIC or newer multi-institutional collections will help evaluate robustness beyond TCGA-BRCA. Second, multimodal fusion represents a promising extension: integrating mammography, ultrasound, MRI, histopathology, and even genomic data may yield more powerful models for risk stratification and treatment planning. A system that jointly reasons over imaging and pathology could better guide decisions such as biopsy, surgery, or systemic therapy. Third, computational efficiency and deployment constraints need to be addressed. Some of the proposed models, especially HMS-T, are heavy in terms of GPU memory and inference time (for example, around 23 GB per WSI). Techniques like pruning, quantization, and knowledge distillation could be used to produce lighter versions that preserve most of the accuracy. Semi-supervised and self-supervised learning on large pools of unlabelled mammograms and slides may further enhance robustness without requiring large annotated datasets. Finally, prospective clinical studies, where radiologists and pathologists work with these tools in real time, will be essential to measure gains in diagnostic accuracy, speed, and user satisfaction, and to uncover practical issues that are not visible in retrospective experiments.

In conclusion, this thesis has presented a comprehensive deep learning framework for improving breast cancer detection, segmentation, staging, and prognosis. Each proposed model—TL-ESMNet, SE-CRNN, MsFuNet, and HMS-T—targets specific gaps in the current landscape and contributes to a more reliable, integrated diagnostic pipeline. Together, they show that, with careful architecture and training design, AI systems can achieve high performance on complex medical imaging tasks while

remaining interpretable and reasonably generalizable. The work advances the technical state of the art but stays closely aligned with clinical needs, bringing these methods closer to real-world adoption. By supporting earlier detection and more accurate diagnosis, such AI tools have the potential to improve patient outcomes through timely and tailored treatment. The results and discussions in this thesis provide a strong foundation for next-generation intelligent breast cancer care systems and indicate several realistic paths to move from laboratory prototypes toward everyday clinical practice.

REFERENCES

[1] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA: a Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, Feb. 2021, doi: https://doi.org/10.3322/caac.21660.

[2] Abdulkerim Capar *et al.*, "An interpretable framework for inter-observer agreement measurements in TILs scoring on histopathological breast images: A proof-of-principle study," *PLoS ONE*, vol. 19, no. 12, pp. e0314450–e0314450, Dec. 2024, doi: https://doi.org/10.1371/journal.pone.0314450.

[3] C.-W. Kao, C.-J. Chiang, and W.-C. Lee, "Comparative Effectiveness of the Revised American Joint Committee on Cancer Staging System," *American Journal of Epidemiology*, Aug. 2024, doi: https://doi.org/10.1093/aje/kwae333.

[4] W. Gómez-Flores and W. Coelho de Albuquerque Pereira, "A comparative study of pre-trained convolutional neural networks for semantic segmentation of breast tumours in ultrasound," *Computers in Biology and Medicine*, vol. 126, p. 104036, Nov. 2020, doi: https://doi.org/10.1016/j.compbiomed.2020.104036.

[5] G. Ayana, E. Lee, and S. Choe, "Abstract 5437: Vision transformers for breast cancer human epidermal growth factor receptor 2 (HER2) expression staging without immunohistochemical (IHC) staining," *Cancer Research*, vol. 83, no. 7_Supplement, pp. 5437–5437, Apr. 2023, doi: https://doi.org/10.1158/1538-7445.am2023-5437.

[6] L. G. Falconi, M. Perez, W. G. Aguila, and A. Conci, "Transfer Learning and Fine Tuning in Breast Mammogram Abnormalities Classification on CBIS-DDSM Database," *Advances in Science, Technology and Engineering Systems Journal*, vol. 5, no. 2, pp. 154–165, 2020, doi: https://doi.org/10.25046/aj050220.

[7] Suckling, J., Parker, J., Dance, D., Astley, S., Hutt, I., Boggis, C., Ricketts, I., Stamatakis, E., Cerneaz, N., Kok, S., Taylor, P., Betal, D., & Savage, J. (2015). Mammographic Image Analysis Society (MIAS) database v1.21. Apollo - University of Cambridge Repository. https://doi.org/10.17863/CAM.105113

[8] Gao Yuan-fu, J. Lin, Y. Zhou, and R. Lin, "The application of traditional machine learning and deep learning techniques in mammography: a review," *Frontiers in Oncology*, vol. 13, Aug. 2023, doi: https://doi.org/10.3389/fonc.2023.1213045.

[9] X.-X. Yin, L. Sun, Y. Fu, R. Lu, and Y. Zhang, "U-Net-Based Medical Image

Segmentation," *Journal of Healthcare Engineering*, vol. 2022, p. 4189781, Apr. 2022, doi: https://doi.org/10.1155/2022/4189781.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *Lecture Notes in Computer Science*, vol. 9351, pp. 234–241, 2015, doi: https://doi.org/10.1007/978-3-319-24574-4_28.

[11] Z. Huang, X. Zhu, M. Ding, and X. Zhang, "Medical Image Classification Using a Light-Weighted Hybrid Neural Network Based on PCANet and DenseNet," *IEEE Access*, vol. 8, pp. 24697–24712, 2020, doi: https://doi.org/10.1109/access.2020.2971225.

[12] A. Ben Hamida *et al.*, "Deep learning for colon cancer histopathological images analysis," *Computers in Biology and Medicine*, vol. 136, p. 104730, Sep. 2021, doi: https://doi.org/10.1016/j.compbiomed.2021.104730.

[13] J. Yao, X. Zhu, J. Jonnagaddala, N. Hawkins, and J. Huang, "Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks," *Medical Image Analysis*, vol. 65, p. 101789, Oct. 2020, doi: https://doi.org/10.1016/j.media.2020.101789.

[14] D. W. Podlich and M. Cooper, "QU-GENE: a simulation platform for quantitative analysis of genetic models," *Bioinformatics*, vol. 14, no. 7, pp. 632–653, Aug. 1998, doi: https://doi.org/10.1093/bioinformatics/14.7.632.

[15] Meshrif Alruily, A. A. Mahmoud, Hisham Allahem, A. M. Mostafa, Hosameldeen Shabana, and M. Ezz, "Enhancing Breast Cancer Detection in Ultrasound Images: An Innovative Approach Using Progressive Fine-Tuning of Vision Transformer Models," *International Journal of Intelligent Systems*, vol. 2024, no. 1, Jan. 2024, doi: https://doi.org/10.1155/int/6528752.

[16] N. Shrivastava and J. Bharti, "Breast Tumour Detection in Digital Mammogram Based on Efficient Seed Region Growing Segmentation," *IETE Journal of Research*, pp. 1–13, Jan. 2020, doi: https://doi.org/10.1080/03772063.2019.1710583.

[17] K. Mridha, T. Sarker, S. D. Bappon, and S. M. Sabuj, "Attention U-Net: A Deep Learning Approach for Breast Cancer Segmentation," *2023 International Conference on Quantum Technologies, Communications, Computing, Hardware and Embedded Systems Security (iQ-CCHESS)*, pp. 1–6, Sep. 2023, doi: https://doi.org/10.1109/iq-cchess56596.2023.10391674.

[18] H. Zhang, J. Xu, M. Wang, and Y. Li, "DenseATT-Net: Densely-Connected Neural Network with Intensive Attention Modules for 3D ABUS Mass

Segmentation," *2022 7th International Conference on Image, Vision and Computing (ICIVC)*, vol. 28, pp. 348–353, Jul. 2022, doi: https://doi.org/10.1109/icivc55077.2022.9886080.

[19] S. Lu, S. Wang, and Y. Zhang, "SAFNet: A deep spatial attention network with classifier fusion for breast cancer detection," *Computers in Biology and Medicine*, vol. 148, pp. 105812–105812, Sep. 2022, doi: https://doi.org/10.1016/j.compbiomed.2022.105812.

[20] Y. S. Leong, K. Hasikin, K. W. Lai, N. Mohd Zain, and M. M. Azizan, "Microcalcification Discrimination in Mammography Using Deep Convolutional Neural Network: Towards Rapid and Early Breast Cancer Diagnosis," *Frontiers in Public Health*, vol. 10, Apr. 2022, doi: https://doi.org/10.3389/fpubh.2022.875305.

[21] G. E. Park, S. H. Kim, Y. Nam, J. Kang, M. Park, and B. J. Kang, "3D Breast Cancer Segmentation in DCE-MRI Using Deep Learning With Weak Annotation," *Journal of Magnetic Resonance Imaging*, vol. 59, no. 6, pp. 2252–2262, Aug. 2023, doi: https://doi.org/10.1002/jmri.28960.

[22] W. Ding, H. Zhang, S. Zhuang, Z. Zhuang, and Z. Gao, "Multi-view stereoscopic attention network for 3D tumour classification in automated breast ultrasound," *Expert Systems with Applications*, vol. 234, p. 120969, Dec. 2023, doi: https://doi.org/10.1016/j.eswa.2023.120969.

[23] J. Peng, C. Bao, C. Hu, X. Wang, W. Jian, and W. Liu, "Automated mammographic mass detection using deformable convolution and multiscale features," *Medical & Biological Engineering & Computing*, vol. 58, no. 7, pp. 1405–1417, Apr. 2020, doi: https://doi.org/10.1007/s11517-020-02170-4.

[24] W. M. Salama and M. H. Aly, "Deep learning in mammography images segmentation and classification: Automated CNN approach," *Alexandria Engineering Journal*, vol. 60, no. 5, pp. 4701–4709, Oct. 2021, doi: https://doi.org/10.1016/j.aej.2021.03.048.

[25] S. Poudel and S.-W. Lee, "Deep multi-scale attentional features for medical image segmentation," *Applied Soft Computing*, vol. 109, p. 107445, Sep. 2021, doi: https://doi.org/10.1016/j.asoc.2021.107445.

[26] R. Gnanasekaran and H. A. Mottola, "Flow injection determination of penicillins using immobilized penicillinase in a single bead string reactor," *Analytical Chemistry*, vol. 57, no. 6, pp. 1005–1009, May 1985, doi: https://doi.org/10.1021/ac00283a009.

[27] X. Wang, G. Liang, Y. Zhang, H. Blanton, Z. Bessinger, and N. Jacobs, “Inconsistent Performance of Deep Learning Models on Mammogram Classification,” *Journal of the American College of Radiology*, vol. 17, no. 6, pp. 796–803, Jun. 2020, doi: https://doi.org/10.1016/j.jacr.2020.01.006.

[28] L. Wang, M. Liu, Y. Wang, X. Bai, M. Zhu, and F. Zhang, “A multi-scale method based on U-Net for brain tumour segmentation,” *2022 7th International Conference on Communication, Image and Signal Processing (CCISP)*, vol. 34, pp. 271–275, Nov. 2022, doi: https://doi.org/10.1109/ccisp55629.2022.9974427.

[29] S. Seo, Y.-G. Suh, D.-W. Kim, G. Kim, J.-W. Han, and B. Han, “Learning to Optimize Domain Specific Normalization for Domain Generalization,” *Lecture Notes in Computer Science*, pp. 68–83, Jan. 2020, doi: https://doi.org/10.1007/978-3-030-58542-6_5.

[30] D. Kim, Y. Yoo, S. Park, J. Kim, and J. Lee, “SelfReg: Self-supervised Contrastive Regularization for Domain Generalization,” *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, doi: https://doi.org/10.1109/iccv48922.2021.00948.

[31] H. Liu, Q. Zhang, and Y. Liu, “Image Segmentation of Bladder Cancer Based on DeepLabv3+,” *Lecture notes in electrical engineering*, pp. 614–621, Oct. 2021, doi: https://doi.org/10.1007/978-981-16-6320-8_62.

[32] T. S. Sheikh, Y. Lee, and M. Cho, “Histopathological Classification of Breast Cancer Images Using a Multi-Scale Input and Multi-Feature Network,” *Cancers*, vol. 12, no. 8, p. 2031, Jul. 2020, doi: https://doi.org/10.3390/cancers12082031.

[33] S. C. Kosaraju, J. Hao, H. M. Koh, and M. Kang, “Deep-Hipo: Multi-scale receptive field deep learning for histopathological image analysis,” *Methods*, vol. 179, pp. 3–13, Jul. 2020, doi: https://doi.org/10.1016/j.ymeth.2020.05.012.

[34] H. U. Khan, B. Raza, A. Waheed, and H. Shah, “MSF-Model: Multi-Scale Feature Fusion-Based Domain Adaptive Model for Breast Cancer Classification of Histopathology Images,” *IEEE Access*, vol. 10, pp. 122530–122547, 2022, doi: https://doi.org/10.1109/access.2022.3223870.

[35] W. Qi, H. C. Wu, and S. C. Chan, “MDF-Net: A Multi-Scale Dynamic Fusion Network for Breast Tumour Segmentation of Ultrasound Images,” *IEEE Transactions on Image Processing*, vol. 32, pp. 4842–4855, 2023, doi: https://doi.org/10.1109/tip.2023.3304518.

[36] M. Dabass, S. Vashisth, and R. Vig, “A convolution neural network with multi-

level convolutional and attention learning for classification of cancer grades and tissue structures in colon histopathological images," *Computers in Biology and Medicine*, vol. 147, p. 105680, Aug. 2022, doi: https://doi.org/10.1016/j.compbiomed.2022.105680.

[37] Ikram Maouche, Labib Sadek Terrissa, K. Benmohammed, and Noureddine Zerhouni, "An Explainable AI Approach for Breast Cancer Metastasis Prediction Based on Clinicopathological Data," *IEEE transactions on bio-medical engineering/IEEE transactions on biomedical engineering*, vol. 70, no. 12, pp. 3321–3329, Dec. 2023, doi: https://doi.org/10.1109/tbme.2023.3282840.

[38] B. Song *et al.*, "Interpretable and Reliable Oral Cancer Classifier with Attention Mechanism and Expert Knowledge Embedding via Attention Map," *Cancers*, vol. 15, no. 5, p. 1421, Feb. 2023, doi: https://doi.org/10.3390/cancers15051421.

[39] Y-S. Lee *et al.*, "The direct comparison between 7th AJCC staging system and 8th AJCC staging system for prediction of survival with Korean multicenter HCC patients," *Journal of Hepatology*, vol. 68, pp. S434–S435, Apr. 2018, doi: https://doi.org/10.1016/s0168-8278(18)31108-5.

[40] P. S. Ginter *et al.*, "Histologic grading of breast carcinoma: a multi-institution study of interobserver variation using virtual microscopy," *Modern Pathology*, pp. 1–9, Oct. 2020, doi: https://doi.org/10.1038/s41379-020-00698-2.

[41] António José Avelãs Nuñes, "As duas últimas máscaras do estado capitalista. Doi: 10.5020/2317-2150.2011.v16n2p409," *Pensar*, vol. 16, no. 2, pp. 409–476, Jun. 2012, doi: https://doi.org/10.5020/23172150.2012.409-476.

[42] X. Qu *et al.*, "A VGG attention vision transformer network for benign and malignant classification of breast ultrasound images," *Medical Physics*, Jul. 2022, doi: https://doi.org/10.1002/mp.15852.

[43] J. Paul *et al.*, "Survival outcome prediction of breast carcinomas on whole-slide histopathology images using deep learning.," *Journal of Clinical Oncology*, vol. 42, no. 16_suppl, pp. 1070–1070, Jun. 2024, doi: https://doi.org/10.1200/jco.2024.42.16_suppl.1070.

[44] S. C. Wetstein *et al.*, "Deep learning-based breast cancer grading and survival analysis on whole-slide histopathology images," *Scientific Reports*, vol. 12, no. 1, Sep. 2022, doi: https://doi.org/10.1038/s41598-022-19112-9.

[45] Raktim Kumar Mondol, Ewan K.A. Millar, P. H. Graham, L. Browne, Arcot Sowmya, and E. Meijering, "hist2RNA: An Efficient Deep Learning Architecture to

Predict Gene Expression from Breast Cancer Histopathology Images," vol. 15, no. 9, pp. 2569–2569, Apr. 2023, doi: https://doi.org/10.3390/cancers15092569.

[46] E. Heer, A. Harper, N. Escandor, H. Sung, V. McCormack, and M. M. Fidler-Benaoudia, "Global burden and trends in premenopausal and postmenopausal breast cancer: a population-based study," *The Lancet Global Health*, vol. 8, no. 8, pp. e1027–e1037, Aug. 2020, doi: https://doi.org/10.1016/s2214-109x(20)30215-1.

[47] N. Saffari *et al.*, "Fully Automated Breast Density Segmentation and Classification Using Deep Learning," *Diagnostics*, vol. 10, no. 11, p. 988, Nov. 2020, doi: https://doi.org/10.3390/diagnostics10110988.

[48] J. Bai, R. Posner, T. Wang, C. Yang, and S. Nabavi, "Applying deep learning in digital breast tomosynthesis for automatic breast cancer detection: A review," *Medical Image Analysis*, vol. 71, p. 102049, Jul. 2021, doi: https://doi.org/10.1016/j.media.2021.102049.

[49] V. Romeo *et al.*, "Tumour segmentation analysis at different post-contrast time points: A possible source of variability of quantitative DCE-MRI parameters in locally advanced breast cancer," *European Journal of Radiology*, vol. 126, pp. 108907–108907, May 2020, doi: https://doi.org/10.1016/j.ejrad.2020.108907.

[50] N. G. Anderson, N. L. Anderson, and S. L. Tollaksen, "Proteins of human urine. I. Concentration and analysis by two-dimensional electrophoresis.," *Clinical Chemistry*, vol. 25, no. 7, pp. 1199–1210, Jul. 1979, doi: https://doi.org/10.1093/clinchem/25.7.1199.

[51] X. Qu, J. Brame, Q. Li, and P. J. J. Alvarez, "Nanotechnology for a Safe and Sustainable Water Supply: Enabling Integrated Water Treatment and Reuse," *Accounts of Chemical Research*, vol. 46, no. 3, pp. 834–843, Jun. 2012, doi: https://doi.org/10.1021/ar300029v.

[52] S. Kang, Y. Kang, and S. Tan, "Exploring and Exploiting Multi-Modality Uncertainty for Tumour Segmentation on PET/CT," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 9, pp. 5435–5446, Sep. 2024, doi: https://doi.org/10.1109/jbhi.2024.3397332.

[53] N. S. Punn and S. Agarwal, "RCA-IUnet: a residual cross-spatial attention-guided inception U-Net model for tumour segmentation in breast ultrasound imaging," *Machine Vision and Applications*, vol. 33, no. 2, Feb. 2022, doi: https://doi.org/10.1007/s00138-022-01280-3.

[54] Heba Abdel-Nabi, Arafat Awajan, and M. Ali, "A Novel Ensemble Strategy

With Enhanced Cross Attention Encoder-Decoder Framework for Tumour Segmentation in Whole Slide Images," Jun. 2022, doi: https://doi.org/10.1109/icics55353.2022.9811163.

[55] S L Parvathi Sitanaboina, Satyanarayana Reddy Beeram, Harikiran Jonnadula, and Lakshmikanth Paleti, "Attention 3D-CU-Net: Enhancing Kidney Tumour Segmentation Accuracy through Selective Feature Emphasis," *IEEE Access*, vol. 11, pp. 139798–139810, Jan. 2023, doi: https://doi.org/10.1109/access.2023.3340912.

[56] Payam Zarbakhsh, "Spatial Attention Mechanism and Cascade Feature Extraction in a U-Net Model for Enhancing Breast Tumour Segmentation," *Applied sciences*, vol. 13, no. 15, pp. 8758–8758, Jul. 2023, doi: https://doi.org/10.3390/app13158758.

[57] Y. Bi, "The Study of Transfer Learning Effectiveness on Medical Brain Tumour Images," *Applied and Computational Engineering*, vol. 8, no. 1, pp. 401–410, Aug. 2023, doi: https://doi.org/10.54254/2755-2721/8/20230197.

[58] Z. Li *et al.*, "Domain Generalization for Mammography Detection via Multi-style and Multi-view Contrastive Learning," *Lecture notes in computer science*, pp. 98–108, Jan. 2021, doi: https://doi.org/10.1007/978-3-030-87234-2_10.

[59]Y. Zhang, X. Zhang, and W. Zhu, "ANC: Attention Network for COVID-19 Explainable Diagnosis Based on Convolutional Block Attention Module," *Computer Modelling in Engineering & Sciences*, vol. 127, no. 3, pp. 1037–1058, 2021, doi: https://doi.org/10.32604/cmes.2021.015807.

[60] Y. Zhao *et al.*, "A Fast 2-D Otsu lung tissue image segmentation algorithm based on improved PSO," *Microprocessors and Microsystems*, vol. 80, p. 103527, Feb. 2021, doi: https://doi.org/10.1016/j.micpro.2020.103527.

[61] Y. Zhang, Y. Li, and D. He, "U-Shaped Network with Multi-Scale Feature Extraction for Lesion Region Segmentation in Gastric Cancer Images," vol. 71, pp. 234–238, May 2023, doi: https://doi.org/10.1109/iciba56860.2023.10165314.

[62] Q. Gao and M. Almekkawy, "ASU-Net++: A nested U-Net with adaptive feature extractions for liver tumour segmentation," *Computers in Biology and Medicine*, vol. 136, p. 104688, Sep. 2021, doi: https://doi.org/10.1016/j.compbiomed.2021.104688.

[63] R. Zou *et al.*, "An Interactive Image Segmentation Method Based on Multi-Level Semantic Fusion," *Sensors*, vol. 23, no. 14, pp. 6394–6394, Jul. 2023, doi: https://doi.org/10.3390/s23146394.

[64] F. Huang and Z. Li, “Study of Two-level P2P Model on Self-adaptive Dynamic Network,” *Modern Applied Science*, vol. 4, no. 7, Jun. 2010, doi: https://doi.org/10.5539/mas.v4n7p59.

[65] Y. Ding *et al.*, “MVFusFra: A Multi-View Dynamic Fusion Framework for Multimodal Brain Tumour Segmentation,” *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 4, pp. 1570–1581, Apr. 2022, doi: https://doi.org/10.1109/jbhi.2021.3122328.

[66] Balamurugan A.G, S. Srinivasan, Preethi D, M. P, Sandeep Kumar Mathivanan, and Mohd Asif Shah, “Robust brain tumour classification by fusion of deep learning and channel-wise attention mode approach,” *BMC medical imaging*, vol. 24, no. 1, Jun. 2024, doi: https://doi.org/10.1186/s12880-024-01323-3.

[67] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, “INbreast,” *Academic Radiology*, vol. 19, no. 2, pp. 236–248, Feb. 2012, doi: https://doi.org/10.1016/j.acra.2011.09.014.

[68] D. Müller, I. Soto-Rey, and F. Kramer, “Towards a guideline for evaluation metrics in medical image segmentation,” *BMC Research Notes*, vol. 15, no. 1, Jun. 2022, doi: https://doi.org/10.1186/s13104-022-06096-y.

[69] H. Cao *et al.*, “Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation,” *Lecture Notes in Computer Science*, vol. 13803, pp. 205–218, 2023, doi: https://doi.org/10.1007/978-3-031-25066-8_9.

[70] J. Chen *et al.*, “TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation,” Feb. 2021, doi: https://doi.org/10.48550/arxiv.2102.04306.

[71] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation,” *Nature Methods*, vol. 18, no. 2, pp. 203–211, Dec. 2020, doi: https://doi.org/10.1038/s41592-020-01008-z.

[72] M. Ma *et al.*, “Predicting the molecular subtype of breast cancer and identifying interpretable imaging features using machine learning algorithms,” *European Radiology*, vol. 32, no. 3, pp. 1652–1662, Oct. 2021, doi: https://doi.org/10.1007/s00330-021-08271-4.

[73] C. Militello *et al.*, “Semi-automated and interactive segmentation of contrast-enhancing masses on breast DCE-MRI using spatial fuzzy clustering,” *Biomedical Signal Processing and Control*, vol. 71, pp. 103113–103113, Jan. 2022, doi: https://doi.org/10.1016/j.bspc.2021.103113.

[74] D. G. V and Megha Bhairagond, “Breast Cancer Detection and Prediction using Image Processing and ML,” *International Journal for Research in Applied Science and Engineering Technology*, vol. 11, no. 8, pp. 1902–1911, Aug. 2023, doi: https://doi.org/10.22214/ijraset.2023.55503.

[75] M. I. Tsarouchi, A. Hoxhaj, and R. M. Mann, “New Approaches and Recommendations for Risk-Adapted Breast Cancer Screening,” *Journal of Magnetic Resonance Imaging*, Apr. 2023, doi: https://doi.org/10.1002/jmri.28731.

[76] Eman Justaniah, Areej Alhothali, and Ghadah Aldabbagh, “Mammogram Segmentation Techniques: A Review,” *International journal of advanced computer science and applications/International journal of advanced computer science & applications*, vol. 12, no. 5, Jan. 2021, doi: https://doi.org/10.14569/ijacsa.2021.0120564.

[77] A. Al Qurri and M. Almekkawy, “Improved UNet with Attention for Medical Image Segmentation,” *Sensors (Basel, Switzerland)*, vol. 23, no. 20, p. 8589, Oct. 2023, doi: https://doi.org/10.3390/s23208589.

[78] M. W. Sabir *et al.*, “Segmentation of Liver Tumour in CT Scan Using ResU-Net,” *Applied Sciences*, vol. 12, no. 17, p. 8650, Aug. 2022, doi: https://doi.org/10.3390/app12178650.

[79] X. Du *et al.*, “Segmentation and visualization of left atrium through a unified deep learning framework,” *International Journal of Computer Assisted Radiology and Surgery*, vol. 15, no. 4, pp. 589–600, Feb. 2020, doi: https://doi.org/10.1007/s11548-020-02128-9.

[80] J. Cui, L. Wang, and S. Jiang, “A Multi-Scale Cross-Fusion Medical Image Segmentation Network Based on Dual-Attention Mechanism Transformer,” *Applied Sciences*, vol. 13, no. 19, pp. 10881–10881, Sep. 2023, doi: https://doi.org/10.3390/app131910881.

[81] H. Wang, X. Bai, S. Wang, J. Tan, and C. Liu, “Generalization on Unseen Domains via Model-Agnostic Learning for Intelligent Fault Diagnosis,” *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, 2022, doi: https://doi.org/10.1109/tim.2022.3152316.

[82] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature Pyramid Networks for Object Detection,” *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, doi: https://doi.org/10.1109/cvpr.2017.106.

[83] K. LIU, Q. CHEN, C. KANG, W. SU, and G. ZHONG, "Optimal operation strategy for distributed battery aggregator providing energy and ancillary services," *Journal of Modern Power Systems and Clean Energy*, vol. 6, no. 4, pp. 722–732, Jan. 2018, doi: https://doi.org/10.1007/s40565-017-0325-9.

[84] G. Jignesh Chowdary and Pratheepan Yoagarajah, "EU-Net: Enhanced U-shaped Network for Breast Mass Segmentation," *IEEE journal of biomedical and health informatics*, pp. 1–11, Jan. 2024, doi: https://doi.org/10.1109/jbhi.2023.3266740.

[85] M. Tan and Q. V. Le, "EfficientNetV2: Smaller Models and Faster Training," *arxiv.org*, Apr. 2021, doi: https://doi.org/10.48550/arXiv.2104.00298.

[86] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision Transformers for Dense Prediction," *IEEE Xplore*, Oct. 01, 2021. https://ieeexplore.ieee.org/document/9711226

[87] X. Yu, C. Kang, D. S. Guttery, S. Kadry, Y. Chen, and Y.-D. Zhang, "ResNet-SCDA-50 for Breast Abnormality Classification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 1, pp. 94–102, Jan. 2021, doi: https://doi.org/10.1109/tcbb.2020.2986544.

[88] A. Iqbal and M. Sharif, "BTS-ST: Swin transformer network for segmentation and classification of multimodality breast cancer images," *Knowledge Based Systems*, vol. 267, pp. 110393–110393, May 2023, doi: https://doi.org/10.1016/j.knosys.2023.110393.

[89] S. Qamar, P. Ahmad, and L. Shen, "Dense Encoder-Decoder–Based Architecture for Skin Lesion Segmentation," *Cognitive Computation*, vol. 13, no. 2, pp. 583–594, Feb. 2021, doi: https://doi.org/10.1007/s12559-020-09805-6.

[90] Y. Alzahrani and B. Boufama, "Breast Ultrasound Image Segmentation Model Based Residual Encoder," *2021 4th International Conference on Intelligent Robotics and Control Engineering (IRCE)*, pp. 85–89, Sep. 2021, doi: https://doi.org/10.1109/irce53649.2021.9570898.

[91] G. C. Ates, P. Mohan, and E. Celik, "Dual Cross-Attention for Medical Image Segmentation," *arXiv (Cornell University)*, Jan. 2023, doi: https://doi.org/10.48550/arxiv.2303.17696.

[92] R. Yu *et al.*, "CFFNN: Cross Feature Fusion Neural Network for Collaborative Filtering," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2021, doi: https://doi.org/10.1109/tkde.2020.3048788.

[93] L. Alzubaidi *et al.*, "MEFF – A model ensemble feature fusion approach for tackling adversarial attacks in medical imaging," *Intelligent Systems with Applications*, vol. 22, p. 200355, Jun. 2024, doi: https://doi.org/10.1016/j.iswa.2024.200355.

[94] Pratiksha Dilip Nandanwar, V. M. Wadhai, A. S. Chanchlani, and V. M. Thakare, "Analysis of Pixel Intensity Variation by Performing Morphological Operations for Image Segmentation On Cervical Cancer Pap Smear Image," pp. 1–6, Nov. 2021, doi: https://doi.org/10.1109/iccica52458.2021.9697185.

[95] A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic Modelling Using Deep Belief Networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, Jan. 2012, doi: https://doi.org/10.1109/tasl.2011.2109382.

[96] G. E. Harbeck, "A practical field technique for measuring reservoir evaporation utilizing mass-transfer theory," *USGS professional paper*, pp. 101–105, Jan. 1962, doi: https://doi.org/10.3133/pp272e.

[97] J. Ferlay *et al.*, "Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods," *International Journal of Cancer*, vol. 144, no. 8, Dec. 2018, doi: https://doi.org/10.1002/ijc.31937.

[98] Z. Rezaei Asl, G. Sepehri, and M. Salami, "Probiotic treatment improves the impaired spatial cognitive performance and restores synaptic plasticity in an animal model of Alzheimer's disease," *Behavioural Brain Research*, vol. 376, p. 112183, Dec. 2019, doi: https://doi.org/10.1016/j.bbr.2019.112183.

[99] Alan Fuad Jahwar and Adnan Mohsin Abdulazeez, "Segmentation and Classification for Breast Cancer Ultrasound Images Using Deep Learning Techniques: A Review," May 2022, doi: https://doi.org/10.1109/cspa55076.2022.9781824.

[100] P. Yadav and R. Rajput, "Pedicled Transfer of Vascularized Scapular Bone Graft to the Humerus," *Plastic & Reconstructive Surgery*, vol. 107, no. 1, pp. 140–142, Jan. 2001, doi: https://doi.org/10.1097/00006534-200101000-00020.

[101] D. E. Frankhauser *et al.*, "Spatiotemporal strategies to identify aggressive biology in precancerous breast biopsies," *WIREs Mechanisms of Disease*, vol. 13, no. 2, Oct. 2020, doi: https://doi.org/10.1002/wsbm.1506.

[102] F. Shahidi, S. Mohd Daud, H. Abas, N. A. Ahmad, and N. Maarop, "Breast Cancer Classification Using Deep Learning Approaches and Histopathology Image: A Comparison Study," *IEEE Access*, vol. 8, pp. 187531–187552, 2020, doi:

https://doi.org/10.1109/access.2020.3029881.

[103] M. Krestenitis, G. Orfanidis, K. Ioannidis, K. Avgerinakis, S. Vrochidis, and I. Kompatsiaris, "Oil Spill Identification from Satellite Images Using Deep Neural Networks," *Remote Sensing*, vol. 11, no. 15, p. 1762, Jul. 2019, doi: https://doi.org/10.3390/rs11151762.

[104] Q. Niu, P. Ao, and D. J. Thouless, "Niu, Ao, and Thouless Reply":, *Physical Review Letters*, vol. 75, no. 5, pp. 975–975, Jul. 1995, doi: https://doi.org/10.1103/physrevlett.75.975.

[105] M. Gheisari and B. Esmaeili, "Applications and requirements of unmanned aerial systems (UASs) for construction safety," *Safety Science*, vol. 118, pp. 230–240, Oct. 2019, doi: https://doi.org/10.1016/j.ssci.2019.05.015.

[106] H. Moreira, A. Szyjka, and K. Gąsiorowski, "Chemopreventive activity of celastrol in drug–resistant human colon carcinoma cell cultures," *Oncotarget*, vol. 9, no. 30, Mar. 2018, doi: https://doi.org/10.18632/oncotarget.25014.

[107] N. A. Saeed, N. F. Muhammad, N. J. Ahmad, N. H. Akhtar, None Muhammad Adeel, and N. A. Ahmad, "The Role of Early Diagnosis and Intervention in Improving Outcomes for Lung Cancer," *Deleted Journal*, vol. 2, no. 1, pp. 126–132, Jun. 2024, doi: https://doi.org/10.62497/irabcs.2024.54.

[108] H. Zhu and B. E. Doğan, "American Joint Committee on Cancer's Staging System for Breast Cancer, Eighth Edition: Summary for Clinicians," *European Journal of Breast Health*, vol. 17, no. 3, pp. 234–238, Jun. 2021, doi: https://doi.org/10.4274/ejbh.galenos.2021.2021-4-3.

[109] R. Deng *et al.*, "Cross-scale multi-instance learning for pathological image diagnosis," *Medical Image Analysis*, vol. 94, pp. 103124–103124, Feb. 2024, doi: https://doi.org/10.1016/j.media.2024.103124.

[110] E. Gresser *et al.*, "Radiomics Signature Using Manual Versus Automated Segmentation for Lymph Node Staging of Bladder Cancer," *European urology focus*, vol. 9, no. 1, pp. 145–153, Jan. 2023, doi: https://doi.org/10.1016/j.euf.2022.08.015.

[111] H. Elmannai, M. Hamdi, and A. AlGarni, "Deep Learning Models Combining for Breast Cancer Histopathology Image Classification," *International Journal of Computational Intelligence Systems*, vol. 14, no. 1, p. 1003, 2021, doi: https://doi.org/10.2991/ijcis.d.210301.002.

[112] M. Desai and M. Shah, "An anatomization on Breast Cancer Detection and

Diagnosis employing Multi-layer Perceptron Neural Network (MLP) and Convolutional Neural Network (CNN),” *Clinical eHealth*, vol. 4, Nov. 2020, doi: https://doi.org/10.1016/j.ceh.2020.11.002.

[113] K. J. DSOUZA and Z. A. ANSARI, “HISTOPATHOLOGY IMAGE CLASSIFICATION USING HYBRID PARALLEL STRUCTURED DEEP-CNN MODELS,” *Applied Computer Science*, vol. 18, no. 1, pp. 20–36, Mar. 2022, doi: https://doi.org/10.35784/acs-2022-2.

[114] Ehsan Monjezi, Gholamreza Akbarizadeh, and Karim Ansari-Asl, “RI-ViT: A Multi-scale Hybrid Method Based on Vision Transformer for Breast Cancer Detection in Histopathological Images,” *IEEE Access*, pp. 1–1, Jan. 2024, doi: https://doi.org/10.1109/access.2024.3514322.

[115] I. Sechopoulos, J. Teuwen, and R. Mann, “Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: State of the art,” *Seminars in Cancer Biology*, vol. 72, Jun. 2020, doi: https://doi.org/10.1016/j.semcancer.2020.06.002.

[116] Z. Yu, N. Pan, and J. Zhou, “SFFNet: Shallow Feature Fusion Network Based on Detection Framework for Infrared Small Target Detection,” *Remote Sensing*, vol. 16, no. 22, pp. 4160–4160, Nov. 2024, doi: https://doi.org/10.3390/rs16224160.

[117] S. Hong, S. Hong, S. Hong, and S. Hong, “Brain tumour classification in VIT-B/16 based on relative position encoding and residual MLP,” *PloS one*, vol. 19, no. 7, pp. e0298102–e0298102, Jul. 2024, doi: https://doi.org/10.1371/journal.pone.0298102.

[118] J. Yu, H. Wang, and Y. Hao, “Two-BranchTGNet: A Two-Branch Neural Network for Breast Cancer Subtype Classification,” pp. 1245–1250, Oct. 2023, doi: https://doi.org/10.1145/3644116.3644327.

[119] J. Abaluck *et al.*, “The Impact of Community Masking on COVID-19: A Cluster-Randomized Trial in Bangladesh,” *Science*, Sep. 2021, doi: https://doi.org/10.26085/c3fs3c.

[120] R. Cui, L. Liu, Y. Song, G. Ren, X. Hu, and J. Qin, “Multi-scale contextual learning for medical image segmentation via dual distillation,” *Medical Physics*, vol. 52, no. 2, pp. 787–800, Nov. 2024, doi: https://doi.org/10.1002/mp.17506.

[121] H. M. Balaha, A. E.-S. Hassan, R. A. Ahmed, and M. H. Balaha, “Advancing eye disease detection: A comprehensive study on computer-aided diagnosis with vision transformers and SHAP explainability techniques,” *Biocybernetics and*

Biomedical Engineering, vol. 45, no. 1, pp. 23–33, Jan. 2025, doi: https://doi.org/10.1016/j.bbe.2024.11.005.

[122] Adarsh Sidda, L. R. Biglow, G. Manu, M. Abdallah, and M. R.B, "Importance of accurate clinical staging in patients with early stage HER2+ and triple negative breast cancer.," *Journal of Clinical Oncology*, vol. 40, no. 16_suppl, pp. e12631–e12631, Jun. 2022, doi: https://doi.org/10.1200/jco.2022.40.16_suppl.e12631.

Brief Bio-data of Candidate

Name of Candidate	:	Satyanarayana Reddy Beram
Date of Birth	:	02/05/1976
Contact	:	9963585097
		mzu22007898@mzu.edu.in
Permanent Address	:	S/O: B Konda Reddy,
		Flat 201,
		Kallam Homes Samskruthi Apartments,
		Railway Station road, Nallapadu,
		Guntur, Andhra Pradesh – 522005, India
Married	:	Yes

Educational Details

MCA	:	Osmania University, Hyderabad, with 76%
M.Tech	:	Computer Science and Engineering, JNTUK, Kakinada, Andhra Pradesh with 78%
Ph.D Course work	:	Mizoram University, SGPA of 9.29

List of Publications

Referred Journals (ESCI and SCOPUS):

1. Satyanarayana Reddy Beram, R. Lalchhanhima, and Ksh. Robert Singh, "Enhancing Breast Tumour Segmentation with TL-ESMNet: A Transfer Learning and Ensemble-Based Approach for Mammograms", *Journal of Advances in Information Technology*, Vol. 16, No. 6, pp. 838-853, 2025. doi: 10.12720/jait.16.6.838-853

2. Beram, S.R., Lalchhanhima, R. & Robert Singh, K. MsFuNet: a novel deep learning framework for robust mammogram analysis using context-aware feature pyramid networks and cross-attention feature fusion. *Evolving Systems* **17**, 58 (2026). https://doi.org/10.1007/s12530-026-09825-x

3. Satyanarayana Reddy Beram, R. Lalchhanhima, and Ksh. Robert Singh, "Hierarchical Multi-scale Transformer for Breast Tumour Staging with Visual Interpretability and Risk Stratification," Journal of Image and Graphics, Vol. 14, No. 2, pp. 208-229, 2026.

International Conferences:

1. Beram, S.R., Lalchhanhima, R., Singh, K.R. (2025). A Novel Deep Learning Architecture for Breast Cancer Segmentation and Classification Using Spatial and Temporal Features. In: Lanka, S., Cabezuelo, A.S., Vuppalapati, C. (eds) Trends in Sustainable Computing and Machine Intelligence. ICTSM 2024. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-96-1452-3_14

2. Beram, S.R., Lalchhanhima, R., Singh, K.R. (2026). HMST-Lite: A Lightweight Multi-scale Transformer for Early Breast Cancer Stage Classification. In: Karsh, R.K., Murugan, R., Laskar, R.H., Schuller, B.W. (eds) Advances on Signal Processing and Computer Vision. SIPCOV 2025. Communications in Computer and Information Science, vol 2848. Springer, Cham. https://doi.org/10.1007/978-3-032-15809-3_6

Particulars of the Candidate

Name of Candidate	:	Satyanarayana Reddy Beram
Degree	:	Ph.D.
Department	:	Information Technology
Title of Thesis	:	Analysis of Medical Images of Breast for the Presence of Tumours using Deep Learning Techniques
Date of Admission	:	18th AUGUST, 2022
Approval of Research Proposal	:	
1. DRC	:	9th MAY 2023
2. BOS	:	11th MAY 2023
3. School Board	:	31st MAY 2023
MZU Regn No	:	2300002
Ph.D Regn No	:	MZU/Ph.D./ 2212 of 18.08.2022
Extension	:	No

Head

Department of Information Technology

ABSTRACT

ANALYSIS OF MEDICAL IMAGES OF BREAST FOR THE PRESENCE OF TUMORS USING DEEP LEARNING TECHNIQUES

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

SATYANARAYANA REDDY BERAM

MZU REGN NO: 2300002

Ph.D. REGN NO: MZU/Ph.D./2212 of 18.08.2022

DEPARTMENT OF INFORMATION TECHNOLOGY

SCHOOL OF ENGINEERING AND TECHNOLOGY

FEBRUARY, 2026

ANALYSIS OF MEDICAL IMAGES OF BREAST FOR THE PRESENCE OF TUMORS USING DEEP LEARNING TECHNIQUES

BY

SATYANARAYANA REDDY BERAM

Department of Information Technology

Supervisor: Dr. R LALCHHANHIMA

Joint Supervisor: Dr. KSH ROBERT SINGH

Submitted

In partial fulfillment of the requirement of the Degree of Doctor of Philosophy in Information Technology of Mizoram University, Aizawl.

Abstract: Breast cancer remains a major global health problem, with more than 2.3 million new cases every year. Today, mammography is the main screening tool for early tumor detection, while histopathology of biopsy tissue is the gold standard for final diagnosis and staging. However, these two pillars of diagnosis have clear limitations. Mammograms can miss small or low-contrast lesions, especially in dense breast tissue, and overlapping structures that can leads to both false negatives and false positives. Reading whole-slide histology images is a time-consuming process and subject to the inter-observer variability, because of different pathologists may assign different grades and stages to the same case. Traditional computer-aided diagnosis systems and early deep learning models have not fully solved these problems: they often overfit to specific datasets, generalize poorly across imaging sources, struggle with multi-scale context, and provide limited interpretability for clinical users. Together, these issues create a strong need for advanced computational frameworks that improve accuracy, robustness, and consistency from detection through to prognosis.

This thesis introduces and evaluates three related deep learning frameworks—MsFuNet, TL-ESMNet, SE-CRNN and HMS-T—designed to address these challenges along the breast cancer diagnostic pathway. Taken together, they cover (i) tumor detection and segmentation in mammograms, (ii) benign-versus-malignant classification, (iii) tumor staging from histopathology, and (iv) preliminary survival risk prediction. MsFuNet (Multi-scale Fusion Network) is a mammogram analysis framework that performs lesion segmentation and malignancy classification in a single model. It uses a context-aware convolutional feature pyramid network to capture multi-scale visual cues (i.e. from tiny microcalcifications to large masses) so that fine local details are combined with global breast structure. A cross-attention fusion module merges features across scales and highlights both the local and contextual signals, which helps to refine tumor boundaries and improve classification robustness. To reduce overfitting and improve generalization across different mammography datasets (DDSM, INbreast, MIAS), MsFuNet employs adversarial domain-agnostic regularization and extensive data augmentation. By

jointly optimizing segmentation and classification, it also simplifies the workflow and reduces computational cost compared with using separate models.

TL-ESMNet (Transfer Learning–Enhanced Segmentation Model) focuses more specifically on tumor segmentation (and downstream classification) in mammograms under limited annotation or small-data conditions. It leverages pretrained CNN encoders such as ResNet or VGG to extract robust feature representations from mammography images, which is especially important when labelled data are scarce or noisy. TL-ESMNet incorporates a dual attention mechanism: spatial attention to emphasize where the tumor is located in the image, and channel attention to emphasize what features are most indicative of tumor tissue. On top of this, a dynamic ensemble fusion strategy integrates outputs from multiple specialized segmentation models. This adaptive fusion, conditioned on tumor characteristics, improves overall segmentation accuracy and sensitivity to subtle, small, or irregularly shaped lesions that standard single models often fail to detect. In this way, TL-ESMNet directly targets one of the persistent weaknesses of mammography by strengthening the detection of small and high-risk tumors.

This thesis also includes a framework called SE-CRNN, which combines the spatial attention and temporal context to improve the breast cancer analysis from mammograms. It has two main parts: SUNet for tumor segmentation and TeResNet for classifying lesions as benign or malignant. SUNet uses spatial excitation to focus on important tumor regions, while TeResNet applies a recurrent layer to learn feature relationships across views or sequences. This helps the model capture more context and reduce false predictions. SE-CRNN showed better performance than existing models on segmentation and classification tasks, especially for difficult cases with dense tissue or subtle lesions.

HMS-T (Hierarchical Multi-Scale Transformer) is designed for the histopathology domain, with the goal of automating tumor staging and providing an initial prognosis. HMS-T processes whole-slide images using a multi-scale Vision Transformer architecture that mimics how pathologists move across magnifications. It uses three ViT branches at 5× (low magnification for tissue architecture), 10×

(medium magnification for glandular structures), and 20× (high magnification for cellular and nuclear detail). A cross-scale attention fusion module then integrates these multi-level representations to enable holistic and accurate AJCC stage classification (Stage I–IV). To support transparency and clinical trust, HMS-T produces attention heatmaps that are quantitatively compared with expert-annotated tumor regions, achieving high spatial agreement. This shows that the model is focusing on medically relevant regions. In addition, HMS-T includes a lightweight survival risk prediction module that combines learned histology features with simple clinical variables (such as age and receptor status) to output a preliminary risk score. This dual output, covering stage and risk, reflects real-world practice, where staging and prognosis are tightly linked, and moves the system beyond pure diagnosis toward outcome prediction.

All three frameworks are evaluated on diverse, real-world datasets. For mammography tasks, the experiments use DDSM, INbreast, and MIAS, which together include film and digital mammograms with varying image quality and annotation detail, allowing us to test generalization across populations and acquisition conditions. For histopathology, the TCGA-BRCA cohort of breast cancer whole-slide images is used to train and validate HMS-T for staging and survival prediction. The results show clear performance gains. MsFuNet achieves consistently high segmentation accuracy, with average tumor Dice scores around 0.88 across mammogram datasets and above 0.90 on the high-quality INbreast set, outperforming several state-of-the-art baselines by around 2–3%. Its average pixel-level accuracy reaches 93.2%, and even on the small MIAS dataset, with limited training samples, MsFuNet maintains a Dice of about 0.86, roughly 2% higher than the next-best method. For benign-versus-malignant classification on DDSM, MsFuNet obtains an F1-score of about 93.7% and an overall accuracy of about 93.5%, outperforming a strong Swin Transformer classifier by approximately 3–4%, and notably reducing false negatives, which is critical in screening.

TL-ESMNet also delivers strong gains in mammography segmentation, with tumor Dice scores in the range 0.88–0.91, clearly higher than popular architectures such as U-Net, TransUNet, and Swin-Unet, which reach roughly 0.77–0.81 in the same

experiments. For example, in one evaluation TL-ESMNet attains a mean Dice of 0.88 compared with 0.81 for the closest Swin-Unet baseline. These improvements are consistent across folds and datasets, indicating robust performance that is not tied to a single data split. Qualitative analysis shows that TL-ESMNet better highlights small or low-contrast tumors that other methods tend to miss, supporting the benefit of its attention and ensemble design. HMS-T achieves high accuracy for automatic tumor staging on TCGA slides, with average staging accuracy around 91.5%, macro F1 $\approx$ 0.89, and a Quadratic Weighted Kappa of about 0.92, all significantly higher than baseline CNN and single-scale transformer models (which remain below 87% accuracy). The integrated survival risk predictor reaches a concordance index of about 0.74, compared with 0.68 for a traditional Cox proportional hazards model using only clinical variables. This suggests that combining multi-scale histology features with clinical data can improve prognostic prediction and support more personalized treatment planning. The attention maps produced by HMS-T also align with key histopathological features of each stage, such as highlighting well-formed glands in Stage I and necrotic or poorly differentiated areas in Stage III, reinforcing that the model's decisions are grounded in meaningful visual evidence.

In total, the proposed frameworks advance the state of the art in deep learning for breast cancer imaging by tackling core challenges of generalizability, sensitivity to small tumors, multi-scale feature learning, and interpretability. By spanning the workflow from screening mammograms to histopathology-based staging and early outcome estimation, MsFuNet, TL-ESMNet, and HMS-T form a cohesive AI-assisted pipeline for breast cancer care. The reported improvements in detection, segmentation, classification, and staging, together with built-in explanation mechanisms, indicate strong clinical relevance. These methods have the potential to reduce missed cancers in screening, standardize and speed up diagnostic workflows, and provide earlier risk estimates to guide therapy. Overall, this thesis shows that carefully designed deep learning frameworks can address real clinical needs and move AI a step closer to routine use in breast cancer management.