成都理工大學中英合作办学
CDUT Sino-British Collaborative Education

OXFORD BROOKES UNIVERSITY

# UNDERGRADUATE PROJECT REPORT

| | |
|---|---|
| **Project Title:** | **Early Cancer Detection Using Multi-Scale CNN and Transformer-Based Deep Learning Approaches** |
| **Surname:** | **Luo** |
| **First Name:** | **Xinyu** |
| **Student Number:** | **202118020329** |
| **Supervisor Name:** | **Grace Ugochi Nneji** |
| **Module Code:** | **CHC 6096** |
| **Module Name:** | **Project** |
| **Date Submitted:** | **May 6th, 2024** |

**Chengdu University of Technology Oxford Brookes College**

**Chengdu University of Technology**

**BSc (Single Honours) Degree Project**

Programme Name: **Software Engineering**

Module No.: **CHC 6096**

Surname: Luo

First Name: Xinyu

Project Title: Early Cancer Detection Using Multi-Scale CNN and Transformer-Based Deep Learning Approaches

Student No.: 202118020329

Supervisor: Grace Ugochi Nneji

2ND Supervisor (if applicable): **Not Applicable**

Date submitted: **May 6th, 2025**

*A report submitted as part of the requirements for the degree of BSc (Hons) in Software Engineering*

*At*

**Chengdu University of Technology Oxford Brookes College**

## Declaration

**Student Conduct Regulations**:

Please ensure you are familiar with the regulations in relation to Academic Integrity. The University takes this issue very seriously and students have been expelled or had their degrees withheld for cheating in assessment. It is important that students having difficulties with their work should seek help from their tutors rather than be tempted to use unfair means to gain marks. Students should not risk losing their degree and undermining all the work they have done towards it. You are expected to have familiarised yourself with these regulations.

https://www.brookes.ac.uk/regulations/current/appeals-complaints-and-conduct/c1-1/

Guidance on the correct use of references can be found on www.brookes.ac.uk/services/library, and also in a handout in the Library.

The full regulations may be accessed online at https://www.brookes.ac.uk/students/sirt/student-conduct/

If you do not understand what any of these terms mean, you should ask your Project Supervisor to clarify them for you.

**I declare that I have read and understood Regulations C1.1.4 of the Regulations governing Academic Misconduct, and that the work I submit is fully in accordance with them**.

Signature: Xinyu Luo Date: April 27th, 2025

REGULATIONS GOVERNING THE DEPOSIT AND USE OF OXFORD BROOKES UNIVERSITY MODULAR PROGRAMME PROJECTS AND DISSERTATIONS

Copies of projects/dissertations, submitted in fulfilment of Modular Programme requirements and achieving marks of 60% or above, shall normally be kept by the Oxford Brookes University Library.

**I agree that this dissertation may be available for reading and photocopying in accordance with the Regulations governing the use of the Oxford Brookes University Library.**

Signature: Xinyu Luo Date: April 27th, 2025

**Acknowledgement**

As this project marks the culmination of my undergraduate journey, this acknowledgment extends beyond the scope of this specific work to encompass the people, experiences, and insights that have shaped my entire four-year academic and personal development.

I would like to express my profound gratitude to my supervisor, Dr. Grace U. Nneji, whose guidance has illuminated my academic journey since our first encounter in my sophomore year. Her unwavering support and meticulously high standards have shaped not only this project but my approach to research and learning. Though her expectations often reached for the stars, her willingness to provide hands-on mentorship demonstrated a commitment to excellence that I aspire to carry forward in all my future endeavors. I am equally indebted to Dr. Joojo Walker, both Module Leader for this project and Level Head for Computer Science and Software Engineering Programmes, whose insights and direction have been invaluable. His thoughtful feedback, technical expertise, and steady guidance throughout the development process have significantly enhanced both this project and my broader understanding of the field. The time he invested in helping me navigate complex challenges reflects a dedication to student success that has truly made a difference in my academic growth.

Life at Songlin Yuan 554 has been a rich mosaic of shared experiences, late-night discussions, and mutual growth alongside five extraordinary roommates: Richard, Kevin, Kerry, Steve, and Tristram. In the sanctuary of our dormitory, amidst laughter and occasional chaos, I discovered connections that would shape my university experience. Though my junior year brought fewer nights within these walls, the bonds we forged transcend physical presence and mere cohabitation; they represent a foundation of companionship that has steadied me through the turbulent seas of university life. Even in my absence, the echoes of our conversations and the warmth of those shared moments continued to guide me through each challenge and celebration.

The serendipitous formation of our chat group "一緒 Impact ですよ" during the DevOps Seminar in 2023 marked the beginning of connections that would define my university experience. What began as academic collaboration evolved into a digital sanctuary where we shared technical frustrations alongside life's absurdities. Our memes, candid critiques, and unfiltered commentary created a vibrant collage of humor that sustained us through demanding semesters. Pia's academic rigor, Kevin's reliable assistance, Victor's programming prowess, Edgar's gaming passion, and Bill's perfectly timed humor—all

transcended ordinary classmate relationships, transforming fleeting digital exchanges into cherished memories. I sincerely hope each member of our unlikely digital family finds their path forward illuminated with opportunity and their futures as bright as the intelligence they've displayed in our conversations.

My heart holds a special place for Yinda and Yukai. Witnessing your relationship has been a beautiful lesson in what a genuine connection can be. The moments we shared throughout 2024 have become treasures I'll carry with me always - from our spontaneous adventures to those quiet evenings of reflection. Your journey together continues to inspire, and I hope your path forward brings you both the same warmth and joy you've brought to my life. I also want to acknowledge Justin, whose presence taught me the boundaries I needed to draw and the voice I needed to find. Through my silences that spoke volumes, and the expectations that subtly revealed the distances between us, I've come to sense the profound gravity those moments may have held. There's now a deeper understanding of how uncertainty might have settled where clarity was longed for and for the unrest gently seeded in those moments, I hold a quiet regret. Yet within that dissonance, I uncovered parts of myself previously hidden from view, a difficult mirror to face, but one that ultimately guided me toward more honest self-expression. For that unintended clarity, and the lessons born from our fracture, I remain deeply grateful.

Football, with its rhythm and raw emotion, and rock music, with its rebellious soul and poetic truth, have been twin beacons illuminating even my darkest days. They remind me that passion and beauty exist alongside struggle, painting the world in vibrant hues when colors seem to fade.

I marvel at the elegant paradox of computer science, how the simplest binary operations of ones and zeros, like mathematical poetry, can orchestrate symphonies of complexity that power our digital world. This discipline has taught me that profound intricacy often emerges from fundamental simplicity, a philosophy I find reflected in many aspects of life.

If there is one thread I carry most quietly, it is to my parents, whose quiet devotion sustained me through both quiet stretches and turbulent crossings. Their steadfast support in presence and unspoken in expectation has made many things possible, not least the space to grow into someone with thoughts and questions of his own. I carry with me both the strength they offered and the imprints they never meant to leave. And in the

process of becoming more fully myself, I hold deep respect for the part they continue to play in who I am, and who I am still becoming.

Finally, I extend gratitude to my own persistent spirit, the resilient flame within that continues to flicker and burn despite winds of doubt and storms of challenge. In a universe of infinite possibility, I find myself here, now, alive with curiosity and hope, ready to contribute my verse to the ongoing dialogue of human knowledge and experience.

**Table of Content**

**List of Figures**

**List of Tables**

**Abstract**
This study proposes a hybrid deep learning framework for early breast cancer detection based on histopathological image analysis. The model integrates Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), leveraging EfficientNetV2 for efficient local feature extraction and ViTs enhanced with Shifted Patch Tokenization (SPT) for global context modeling. The framework was trained and evaluated on the BreakHis dataset comprising over 7,900 images at multiple magnification levels, and further validated on the BACH dataset of 400 high-resolution microscopy images. On the BreakHis dataset, the model achieved a test accuracy of 92.1%, an AUC-ROC of 0.9715, and a sensitivity of 94.6%, while on the BACH dataset, it attained an accuracy of 86.7%, AUC-ROC of 0.8700, and sensitivity of 92.5%. These results demonstrate the model's strong classification performance and generalization capability across different datasets. The proposed approach shows considerable potential for clinical applications that require reliable, high-sensitivity diagnostic support in digital pathology.

***Keywords:*** *Breast Cancer Detection, Deep Learning, Convolutional Neural Networks, Vision Transformer, Medical Image Analysis, Histopathological Images, Shifted Patch Tokenization, EfficientNetV2*

**Abbreviations**

- **CNN**: Convolutional Neural Network
- **ViT**: Vision Transformer
- **AUC**: Area Under the Curve
- **ROC**: Receiver Operating Characteristic
- **F1-Score**: The harmonic mean of Precision and Recall
- **LSA**: Learned-Scale Attention
- **SPT**: Shifted Patch Tokenization
- **BreakHis**: Breast Cancer Histopathological Dataset
- **BACH**: ICIAR2018 Breast Cancer Histology Dataset
- **GAP**: Global Average Pooling
- **MBConv**: Mobile Inverted Bottleneck Convolution
- **SVM**: Support Vector Machine
- **k-NN**: k-Nearest Neighbors
- **FP**: False Positive
- **FN**: False Negative
- **TP**: True Positive
- **TN**: True Negative
- **MLP**: Multilayer Perceptron

**Glossary**

- **Breast Cancer Histopathological Images**: Microscopic images of breast tissue, typically captured during biopsy procedures, used for detecting cancerous tissue through visual analysis. These images contain detailed cellular information, which is key for diagnosing benign or malignant tissue.
- **Convolutional Neural Network (CNN)**: A class of deep neural networks, typically used to analyze visual imagery. CNNs are designed to automatically and adaptively learn spatial hierarchies of features through backpropagation.
- **Vision Transformer (ViT)**: A type of deep learning model designed for image recognition tasks. Unlike traditional CNNs, Vision Transformers apply self-attention mechanisms and treat image patches as tokens, enabling the model to capture long-range dependencies across the image.
- **EfficientNetV2**: A family of convolutional neural network models designed to be highly efficient. EfficientNetV2 improves on its predecessors by introducing a compound scaling method to balance network depth, width, and resolution for optimal performance.
- **Shifted Patch Tokenization (SPT)**: A mechanism that enhances the Vision Transformer by shifting the image into multiple spatial views, generating enriched patch embeddings for better spatial feature learning, especially in tasks requiring fine-grained image details.
- **Learned-Scale Attention (LSA)**: A novel attention mechanism that introduces learnable temperature parameters to adjust the distribution of attention scores, enhancing the model's ability to prioritize important features during learning.
- **Multi-Scale CNN**: A neural network approach that processes data at different magnification levels to capture both local and global features in images. This method is especially useful in medical image analysis for identifying structures at various levels of detail.
- **Receiver Operating Characteristic (ROC)**: A graphical representation of a model's diagnostic ability across different thresholds. It plots the true positive rate against the false positive rate and is commonly used to evaluate binary classification models.

- **Area Under the Curve (AUC)**: A performance metric derived from the ROC curve. AUC measures the overall ability of a classifier to distinguish between classes, with higher values indicating better performance.
- **Global Average Pooling (GAP)**: A technique in deep learning used to reduce the spatial dimensions of feature maps by taking the average of each feature map. This operation helps reduce the number of parameters and computations in the network.
- **Batch Normalization**: A technique to improve the training of deep neural networks by normalizing the inputs of each layer, which helps stabilize and speed up the training process.
- **Dropout**: A regularization method used in neural networks to prevent overfitting. During training, random units are dropped (i.e., set to zero) in each layer, forcing the network to learn more robust features.
- **Hyperparameter Tuning**: The process of optimizing a model's hyperparameters (such as learning rate, batch size, and number of layers) to achieve the best performance.
- **Cross-Entropy Loss**: A loss function commonly used in classification problems. It measures the difference between the predicted probability distribution and the true distribution of labels.
- **Precision**: The ratio of correctly predicted positive observations to the total predicted positives. It answers the question: “What proportion of positive predictions were actually correct?”
- **Recall**: The ratio of correctly predicted positive observations to all observations in the actual class. It answers the question: “What proportion of actual positives was correctly identified?”
- **F1-Score**: The harmonic mean of precision and recall. It is used to measure a model’s accuracy when the class distribution is imbalanced.
- **Confusion Matrix**: A performance measurement tool for classification problems. It helps visualize the performance of a model by showing the true positives, false positives, true negatives, and false negatives.

- **Training/Validation/Test Accuracy**: Metrics used to evaluate a model's performance on different subsets of the dataset: training accuracy on the training data, validation accuracy on validation data, and test accuracy on unseen data.

# Chapter 1 Introduction

## 1.1 Background

Cancer detection remains a critical challenge in modern healthcare, with significant implications for patient outcomes, treatment options, and healthcare resource allocation. Early detection of cancer, particularly breast cancer, has been shown to dramatically improve survival rates and reduce treatment costs. However, current detection methods face numerous limitations that affect their efficacy and reliability. The following sections explore the risk factors associated with breast cancer detection and the technical challenges that researchers face when developing novel detection systems. These challenges have motivated the development of advanced artificial intelligence approaches that aim to overcome the limitations of traditional methods while improving diagnostic accuracy.

### 1.1.1 Risk and Factor

Early detection of cancer remains one of the most critical strategies for improving patient outcomes, with significant improvements in survival rates when cancer is diagnosed at early stages [1]. This is particularly true for breast cancer, which continues to be one of the leading causes of death among women worldwide [2]. The impact of early detection on patient outcomes is substantial, as it enables less aggressive treatment options and significantly improves the chances of recovery.

The histopathological distinction between benign (left) and malignant (right) breast tissue, as shown in Figure 1, illustrates the cellular changes that must be accurately identified for proper diagnosis. These microscopic differences form the basis for both traditional detection methods and emerging deep learning approaches.

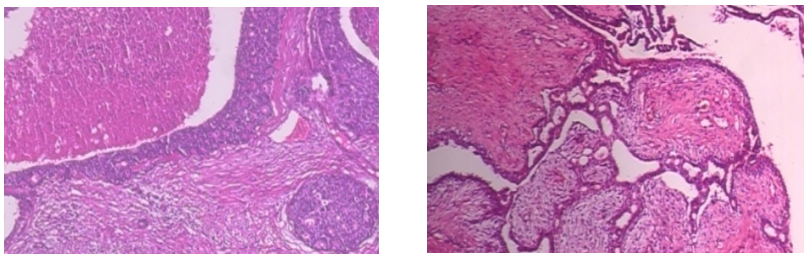

Figure 1 Histopathological comparison of benign (left) and malignant (right)

Traditional cancer detection methods, such as mammography and biopsy, have been instrumental in cancer diagnosis for decades. However, these methods face inherent limitations that affect their reliability and widespread adoption. Mammography, while

effective as a screening tool, has demonstrated concerning rates of both false-positive and false-negative results [3]. According to Ho et al. [4], over a 10-year screening period, the cumulative probability of false-positive results remains a significant concern in breast cancer screening.

The consequences of these diagnostic limitations extend beyond clinical outcomes to encompass economic and psychological impacts. False positives lead to unnecessary follow-up procedures, including additional imaging and biopsies, which increase both healthcare costs and patient anxiety. Conversely, false negatives delay critical treatments that could potentially improve patient outcomes. McGarvey et al. [5] demonstrated a direct correlation between later-stage cancer diagnosis and increased healthcare costs, underscoring the economic implications of delayed detection.

Understanding the multifaceted nature of breast cancer risk is essential for developing comprehensive detection strategies. As illustrated in Figure 2, breast cancer risk factors can be broadly categorized into biological and lifestyle factors. Biological factors include age, family history, genetic predisposition (particularly BRCA1/2 mutations), hormonal influences, breast tissue density, and reproductive history. Lifestyle factors encompass diet and nutrition, physical activity levels, alcohol consumption, smoking habits, environmental exposures, and weight management. These diverse risk factors ultimately contribute to the detection challenges that both traditional and AI-based approaches must address.

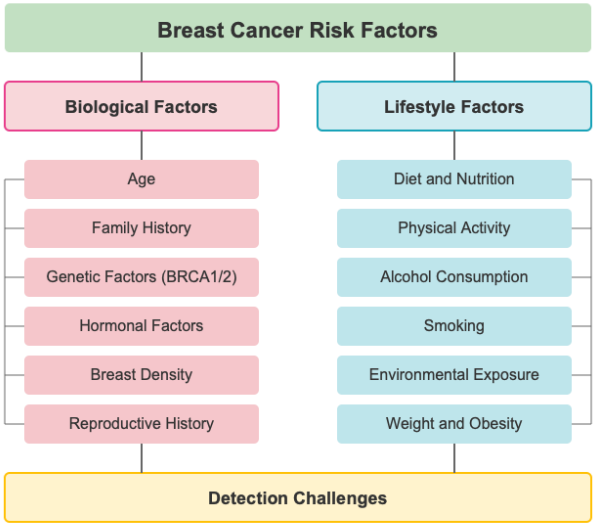


Figure 2 Breast cancer risk factors

Recent advances in artificial intelligence, particularly in deep learning, have shown promising potential for enhancing cancer detection accuracy. Jiang et al. [6] demonstrated that Convolutional Neural Networks (CNNs) have achieved remarkable success in medical image analysis, specifically in identifying patterns within histopathological images. Their ability to learn hierarchical features from input images makes them particularly effective for detecting subtle patterns that might indicate early-stage cancer [7].

### 1.1.2 Challenge

Despite the promising advancements in AI-based cancer detection, significant challenges remain in developing robust and reliable systems for clinical implementation. These challenges span technical, clinical, and implementation domains.

From a technical perspective, CNNs face inherent limitations in their ability to capture comprehensive image context. The hierarchical feature learning capability of CNNs allows them to automatically discover multiple levels of representation, from low-level features like edges and textures to high-level semantic concepts relevant to cancer detection [8]. However, CNNs have restricted receptive fields which limit their ability to capture long-range dependencies and global contextual information. This limitation can lead to suboptimal performance when analyzing complex, high-resolution medical images such as histopathological slides [9].

The emergence of Vision Transformers (ViTs) has introduced new possibilities to address these technical challenges. Originally developed for natural language processing, Transformers have been adapted for computer vision tasks, offering superior capabilities in capturing global dependencies through their self-attention mechanisms [10]. Pereira and Hussain [11] conducted a comprehensive review highlighting how Transformer-based models excel at capturing global context and spatial relationships in medical imaging tasks. This ability to model relationships between distant parts of an image is particularly valuable in cancer detection, where understanding the broader tissue context is often as crucial as identifying local cellular abnormalities.

Current research indicates that combining CNNs with Transformers could potentially address the limitations of each individual approach [12]. As illustrated in Figure 3, the challenges in breast cancer detection can be categorized into three main areas. Technical challenges include image resolution limitations, CNN's limited receptive field, contextual information loss, feature extraction complexity, and long-range dependencies. Clinical

challenges encompass false positives, false negatives, tissue density variation, early-stage detection difficulties, and radiologist interpretation issues. These technical and clinical challenges ultimately lead to implementation challenges related to computational resources, clinical integration, and model interpretability.

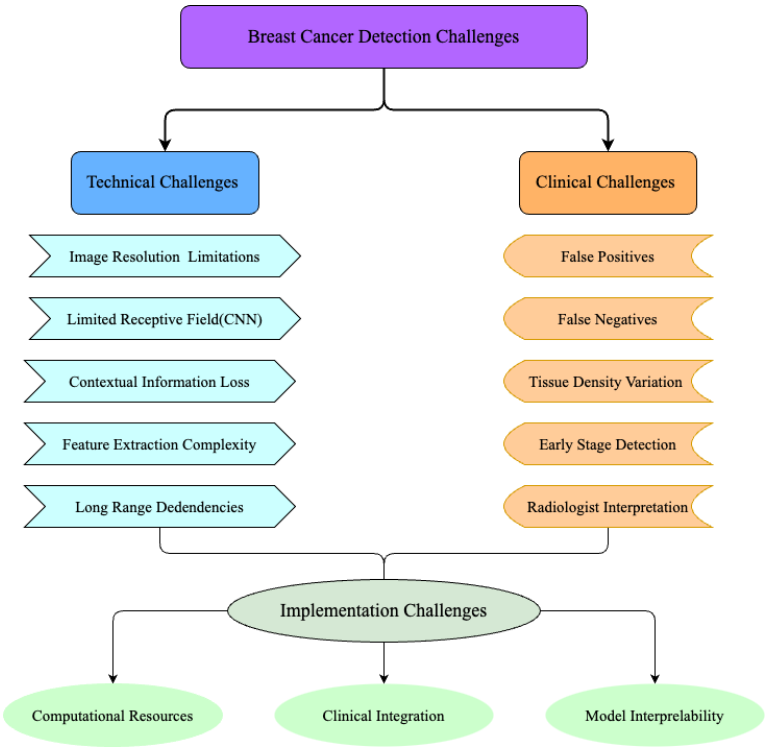


Figure 3 Breast cancer detection challenges

CNNs excel at capturing local features and processing high-resolution images, while Transformers are adept at modelling long-range dependencies and global context [13]. This complementary relationship suggests that a hybrid approach, particularly using multi-scale CNNs combined with Transformer-based architectures, could significantly improve the accuracy and reliability of early cancer detection [14].

From an implementation perspective, the development of robust cancer detection systems requires comprehensive datasets that capture the diversity of clinical presentations. The availability of datasets like BreakHis (Breast Cancer Histopathological Database) [15], which contains microscopic images at various magnification levels (40X, 100X, 200X, and 400X), provides an opportunity to develop and validate hybrid approaches across different scales of tissue analysis. This multi-scale nature aligns well with the proposed hybrid architecture's ability to analyze images at different levels of detail.

The primary challenge in this research is to effectively leverage these technological advances to develop a more accurate and reliable system for early cancer detection, reducing the rate of false positives and negatives that currently challenge traditional diagnostic methods. By combining the strengths of both CNNs and Transformers, this project seeks to create a model that can better understand both the fine-grained details and the broader contextual patterns in histopathological images, ultimately contributing to more accurate and earlier cancer diagnoses.

### 1.2 Aim

The primary aim of this project is to develop and implement an innovative hybrid deep learning framework that significantly advances the field of automated breast cancer detection in histopathological images. This framework combines the complementary strengths of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) to create a more accurate, robust, and clinically relevant diagnostic tool.

This approach addresses fundamental limitations in existing methodologies by integrating an enhanced Vision Transformer architecture with efficient CNN backbones. Specifically, this project aims to implement a Shifted Patch Tokenization (SPT) mechanism that substantially improves the ViT model's ability to capture local spatial information while maintaining its global context awareness. By fusing EfficientNetV2's powerful hierarchical feature extraction capabilities with the long-range dependency modelling of Transformers, this research seeks to develop a model that can identify subtle tissue patterns critical for early cancer detection.

Furthermore, this project aims to implement advanced attention mechanisms through Learned-Scale Attention (LSA), which introduces trainable temperature parameters to optimize attention score distribution. This innovation enables more nuanced feature importance weighting, particularly beneficial for distinguishing between the complex and often subtle differences in benign and malignant tissue structures.

The framework is designed to achieve superior performance in breast cancer classification across multiple magnification levels (40x/100x/200x/400x) using the comprehensive BreakHis dataset, while also validating its generalization capability on the BACH dataset. Emphasis is placed on maintaining the interpretability of results, a critical requirement for clinical adoption. Through this research, the project aims to contribute a methodologically rigorous and technically advanced approach to early cancer detection,

striking a balance between high diagnostic accuracy and practical clinical applicability, with the ultimate goal of improving patient outcomes through earlier and more reliable diagnoses.

## 1.3 Objectives

### 1.3.1 Dataset Processing and Enhancement

- Acquire and organize both the BreakHis and BACH breast cancer histopathological datasets. BreakHis contains 7,909 microscopic images captured at four magnification levels (40x, 100x, 200x, and 400x), while BACH offers additional high-quality annotated slides that support cross-dataset validation.
- Implement comprehensive data preprocessing techniques, including pixel value normalization and image resizing (to 160×160 pixels), ensuring consistency across samples.
- Develop data augmentation strategies such as rotation, scaling, translation, and flipping to increase sample diversity and reduce the risk of overfitting.
- Create efficient data pipelines using TensorFlow to process multi-scale images and support parallel handling of both BreakHis and BACH datasets.

### 1.3.2 Model Architecture Development

- Design and implement a baseline Vision Transformer (ViT) architecture with appropriate patch size and embedding dimensions
- Develop an innovative Shifted Patch Tokenization (SPT) mechanism to enhance local feature extraction capabilities
- Create a hybrid architecture that integrates EfficientNetV2-B0 convolutional backbone with the Vision Transformer
- Implement an advanced Learned-Scale Attention (LSA) mechanism with trainable temperature parameters to optimize attention score distribution

### 1.3.3 Training and Optimization Strategy

- Establish a robust training pipeline with comprehensive logging and monitoring of key metrics
- Implement adaptive learning rate scheduling with inverse time decay and early stopping mechanisms
- Develop model checkpointing to save best-performing model states during training
- Optimize hyperparameters including learning rate, dropout rate, and weight decay through systematic experimentation

- Design efficient gradient computation with appropriate batch sizes and optimization techniques

#### 1.3.4 Performance Evaluation Framework

- Develop a comprehensive evaluation system calculating key metrics including accuracy, precision, recall, and F1 scores
- Create visualization tools for ROC curves, confusion matrices, and training progress analysis
- Implement comparative analysis between the hybrid model and baseline architectures
- Analyze model performance across the 100x magnification level with plans to expand to other magnifications

#### 1.3.5 System Integration and Deployment

- Create a flexible configuration system for model parameters and training settings
- Implement model serialization and loading functionalities for practical deployment
- Develop an end-to-end inference pipeline with appropriate pre and post-processing steps
- Establish error handling and logging systems to ensure robust operation

### 1.4 Project Overview

This project is dedicated to developing and implementing an advanced deep learning system for breast cancer detection through histopathological image analysis, with a particular focus on utilizing and enhancing Vision Transformer (ViT) architectures. The motivation stems from the critical need for accurate and reliable automated cancer detection systems in medical diagnostics, where early and accurate detection can significantly impact patient outcomes.

At its core, the project leverages two well-established datasets: BreakHis, a comprehensive collection of breast cancer histopathological images captured at multiple magnification levels (40x, 100x, 200x, and 400x), and BACH, which provides high-resolution annotated images for cross-validation and generalization analysis. This combination supports both multi-scale feature learning and robust evaluation across diverse clinical image sources. The project implements extensive data augmentation techniques and custom data splitting strategies to ensure reliable model training and validation across datasets.

The technical innovation lies in its novel architectural design, which combines the strengths of Vision Transformers with traditional convolutional approaches. The research develops a hybrid model that integrates a modified ViT architecture with EfficientNetV2, enhanced by an innovative Shifted Patch Tokenization (SPT) mechanism. This combination allows for both efficient local feature extraction and comprehensive global context understanding. Additionally, the model incorporates a learned-scale attention mechanism to optimize the feature extraction capabilities.

The training framework incorporates optimization techniques, including adaptive learning rate scheduling, early stopping mechanisms, and advanced regularization methods such as dropout and weight decay. These components work together to ensure efficient and effective model training while preventing overfitting. The comprehensive validation procedures continuously monitor the model's performance and guide the optimization process.

The evaluation system provides multiple performance metrics, including accuracy, precision, recall, and F1-scores. The framework generates ROC curves, AUC scores, and confusion matrices, accompanied by detailed visualization tools for in-depth performance analysis. This comprehensive evaluation framework allows for detailed comparative analysis across different model configurations and magnification levels.

This project contributes meaningfully to the application of deep learning in medical image analysis by integrating recent architectural advances with practical considerations for clinical deployment. It explores the potential of hybrid models in supporting automated diagnostic tools, aiming to balance methodological innovation with accuracy and reliability in cancer detection. Figure 4 illustrates the complete workflow of this project, from data processing through model development and training to the final deployment and application phases.

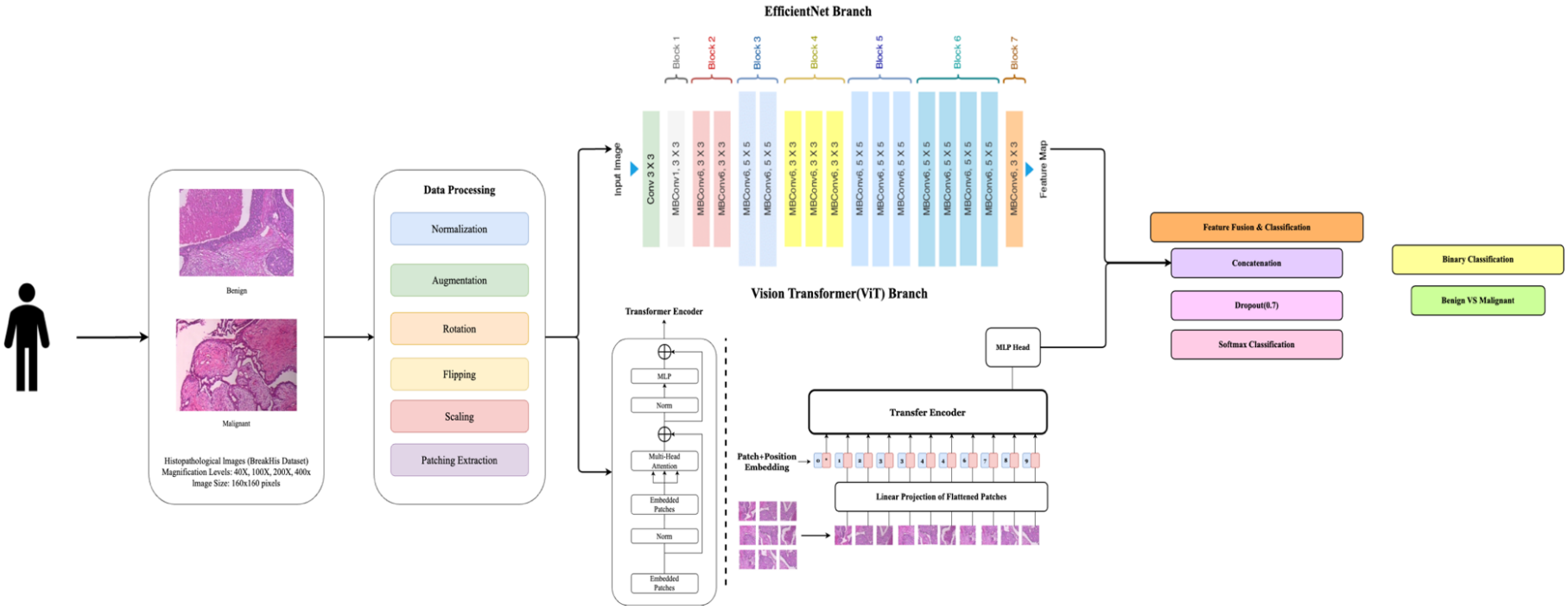


Figure 4 Breast Cancer Detection Project Workflow

### 1.4.1 Scope

The purpose of this study is to develop a hybrid deep learning model combining multi-scale Convolutional Neural Networks (CNNs) and Transformer-based architectures to improve the accuracy and robustness of early cancer detection in histopathological images. The study focuses on leveraging CNNs' ability to capture fine-grained local features and Transformers' capability of modelling long-range dependencies and global context. This hybrid model is expected to outperform traditional deep learning models in detecting cancerous cells, especially in early stages where accurate diagnosis is critical.

The significance of this study lies in its potential to contribute to the field of medical image analysis by enhancing diagnostic precision, reducing false positives and negatives, and ultimately improving patient outcomes. By addressing the limitations of current cancer detection methods, this research can help pave the way for more reliable automated systems in clinical settings, leading to earlier and more accurate cancer diagnoses.

### 1.4.2 Audience

The findings and developments from this deep learning-based research project will benefit several key stakeholder groups in the medical and research communities.

Primary healthcare professionals, particularly pathologists and oncologists, will benefit from the enhanced diagnostic capabilities provided by this automated cancer detection system. The model's ability to analyze histopathological images across multiple magnification levels offers valuable decision support in clinical settings, potentially improving the accuracy and efficiency of cancer diagnosis.

Medical researchers and academic institutions stand to gain from the methodological contributions of this project, particularly the novel integration of Vision Transformers with CNN architectures. The research findings regarding the effectiveness of Shifted Patch Tokenization and learned-scale attention mechanisms provide valuable insights for future developments in medical image analysis.

Healthcare institutions and diagnostic laboratories can utilize this research to enhance their diagnostic workflows. The project's comprehensive evaluation framework and performance metrics offer a blueprint for implementing and assessing similar systems in clinical environments.

Software developers and machine learning engineers in the medical technology sector will find value in the technical implementations and architectural innovations presented in

this research. The project's approach to handling multi-scale medical images and its solutions to common challenges in medical image analysis provide practical insights for similar applications.

Additionally, patients represent an indirect but crucial beneficiary group, as improved diagnostic accuracy and earlier detection capabilities could lead to more timely and appropriate treatment interventions, potentially improving medical outcomes.

# Chapter 2 Background Review

## 2.1 Traditional Methods for Breast Cancer Detection

In the pre-deep learning era, traditional methods for breast cancer detection were predominantly based on manual examination and diagnostic imaging techniques. Mammography, as highlighted by Nelson et al. [3], was a cornerstone of screening, despite its limitations in terms of false positives and negatives. The conventional mammography screening process involves X-ray imaging of the breast tissue, which is then examined by radiologists to identify potential abnormalities. According to Ho et al. [4], over a 10-year screening period, the cumulative probability of false-positive results remains a significant concern, with many women experiencing at least one false-positive recall.

Clinical breast examinations (CBE) [16], performed by healthcare professionals, serve as another traditional approach but suffer from variability in technique and interpretation. Ultrasound imaging provides complementary information to mammography, particularly for dense breast tissue, but also faces challenges in standardization and operator dependency.

## 2.2 Machine Learning Methods for Cancer Detection

Machine learning (ML) has revolutionized oncology research by providing sophisticated tools that can identify subtle patterns in complex medical data, potentially detecting cancer before clinical symptoms manifest. This capability addresses the critical need for early detection, which remains the most effective strategy for improving cancer treatment outcomes.

The integration of ML with advanced assay technologies has enabled the development of multi-cancer early detection (MCED) tests, capable of screening for multiple cancer types simultaneously through minimally invasive procedures. Liquid biopsy analysis, which examines circulating tumour DNA (ctDNA) and other blood-borne biomarkers, has been significantly enhanced by ML algorithms that can detect cancer signals from complex molecular data. Sammut et al. demonstrated this capability by developing a multi-omic ML predictor that integrates diverse biological data to accurately forecast breast cancer therapy responses, illustrating ML's potential to transform diagnostic precision [17].

ML applications span across various cancer types, with particularly promising results in cervical cancer detection. Alsmariy et al. developed an innovative voting classifier that

combines logistic regression, random forest, and decision tree algorithms to achieve superior diagnostic accuracy and sensitivity [18]. In cytology-based screening approaches, ML algorithms analyzing Pap smear images have achieved remarkable accuracy rates, with some models reaching near-perfect classification of cellular abnormalities in controlled settings [19].

Recent research has expanded to explore the complex relationship between the microbiome and cancer. ML models can now effectively analyze the gut microbiome's composition and function to identify signatures associated with cancer development and treatment response. These microbiome-based approaches offer novel biomarkers for early detection and may inform personalized therapeutic strategies [20].

The power of ML in cancer detection lies in its ability to integrate heterogeneous data types, which is genomic, transcriptomic, proteomic, and clinical information into unified predictive frameworks. Advanced ensemble methods and deep learning architectures have demonstrated particular efficacy in handling the complexity inherent in cancer datasets. Current research priorities include developing models that maintain robust performance across diverse populations and clinical environments, addressing the critical challenge of generalizability [21].

As computational capabilities advance and biological datasets expand, ML approaches continue to evolve, offering increasingly sensitive and specific cancer detection methods. Collaborative efforts across disciplines, which bringing together expertise in data science, molecular biology, and clinical oncology, are accelerating progress toward ML-powered cancer screening tools that could significantly improve early detection rates and, ultimately, patient outcomes worldwide.

## 2.3 Deep Learning Methods for Cancer Detection

### 2.3.1 CNN-Based Approaches

In recent studies focusing on breast cancer classification from histopathological images, Convolutional Neural Networks (CNNs) have established themselves as a foundational approach. Albashish et al [22]. leveraged a pre-trained VGG16 model with transfer learning techniques, demonstrating significant improvements in breast cancer classification by integrating CNNs with traditional machine learning classifiers including Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN). This hybrid approach

underscores the potential of combining deep feature extraction capabilities with classical classification methods to enhance diagnostic accuracy.

### 2.3.2 Hybrid CNN-Transformer Models

Hybrid models combining CNNs and Vision Transformers have gained prominence for their complementary strengths. Wang et al.[23] proposed an innovative hybrid CNN-Capsule Network (CapsNet) architecture, where deep feature fusion and enhanced routing mechanisms effectively integrate both convolutional and capsule features to improve classification performance. The dual-channel approach allows the model to capture both local and global image features simultaneously.

Similarly, Abimouloud et al.[24] developed a hybrid Vision Transformer-CNN model that successfully integrates ViT's powerful self-attention mechanism with CNN's efficient feature extraction capabilities. This architectural innovation not only reduced computational costs but also achieved high classification accuracy on the BreakHis dataset, demonstrating the practical benefits of combining these complementary approaches.

### 2.3.3 Vision Transformer Innovations

Vision Transformers (ViTs) have shown exceptional potential in handling complex histopathological image data. Baroni et al.[25] optimized ViT performance through advanced techniques including pretraining on large datasets and implementing sophisticated data augmentation strategies, achieving state-of-the-art results across multiple breast cancer histopathological image datasets.

Gella [26] further demonstrated the effectiveness of fine-tuning a ViT model specifically for the BreakHis dataset, obtaining an unprecedented classification accuracy of 99.99%, which established a new benchmark in the field of breast cancer histopathological image analysis. This remarkable performance highlights the adaptability of transformer architectures to medical imaging tasks when properly optimized.

Alotaibi et al.[27] explored ensemble approaches by combining Vision Transformer (ViT) and Data-Efficient Image Transformer (DeiT) models for breast cancer histopathological image classification. Their research showcased how transformer-based ensemble methods can further enhance classification accuracy and reliability beyond what individual models can achieve.

### 2.3.4 Advanced Learning Frameworks

Multiple instance learning (MIL) has emerged as an effective approach to address the challenges of localizing malignant regions in breast cancer histopathology. Patil et al.[28] introduced an attention-based MIL approach that simultaneously improved both classification accuracy and localization precision, offering enhanced interpretability compared to traditional CNN methods.

Building on transformer architectures, Wang et al.[29] developed an innovative semi-supervised learning framework with Vision Transformers, applying adaptive token sampling techniques to significantly enhance breast cancer classification performance on the BreakHis dataset. This approach demonstrates how modern learning paradigms can be effectively combined with state-of-the-art architectures to address limitations in medical image analysis.

Table 1 Summary Table of Background Review

| Author | Datasets | Methods & Models | Limitation | Results |
|---|---|---|---|---|
| Dheeb Albashish et al. [22] | BreaKHis | VGG16 for feature extraction with classifiers (RBF-SVM, Poly-SVM, KNN, Logistic Regression, NN) | Reliance on pre-trained models; limited to single magnification (40x); high feature dimensionality; lack of misclassification analysis; poor model interpretability. | RBF-SVM achieved 96% accuracy for binary classification, and 89.83% accuracy for multiclass classification at 40x magnification. |
| Pin Wang et al. [23] | BreaKHis | FE-BkCapsNet: a dual-channel network combining CNN and CapsNet features with enhanced routing | Complex dual-channel architecture requires significant computational resources; lack of interpretability in the fusion process; performance | Achieved classification accuracy of 92.71% at 40x, 94.52% at 100x, 94.03% at 200x, and 93.54% at |

| | | | | |
|---|---|---|---|---|
| | | | variations across magnification levels; limited generalizability to other cancer types; insufficient validation on external datasets. | 400x magnifications. |
| Mouhamed Laid Abimoulod et al.[24] | BreaKHis | Hybrid models combining Vision Transformer (ViT), Compact Convolution Transformers (CCT), and Mobile Vision Transformers (MVIT) | Limited cross-dataset validation; high computational requirements for transformer models; inconsistent performance across magnification levels; model complexity may limit clinical implementation; potential overfitting on imbalanced subtype data. | Achieved 98.64% accuracy with ViT, 96.99% with CCT, and 97.52% with MVIT for binary classification at optimal magnifications |
| Giulia Lucrezia Baroni et al. [25] | BACH, BRACS, AIDPATH | Vision Transformer (ViT) pretrained on ImageNet with color normalization and data augmentation | Limited cross-dataset generalization with lower accuracy on BRACS (0.74) compared to BACH (0.91) and AIDPATH (0.92) datasets. | Achieved 0.91 accuracy on BACH, 0.74 on BRACS, and 0.92 on AIDPATH dataset for tumor classification |
| Venkat Gella[26] | BreaKHis | Fine-tuned Vision Transformer (ViT) with Ranger optimizer | Limited dataset validation beyond BreakHis; extreme accuracy claims (99.99%) raise | Achieved an accuracy of 99.99%, precision of 99.98%, and |

| | | | | |
|---|---|---|---|---|
| | | | concerns about potential overfitting despite cross-validation. | recall of 99.99% for binary classification |
| Amira Alotaibi et al.[27] | BreakHis | Ensemble model of Vision Transformer (ViT) and Data-Efficient Image Transformer (DeiT) | Limited architectural novelty; ensemble approach increases computational complexity; model has not been tested on external datasets for generalization assessment. | Achieved 98.17% accuracy, 98.18% precision, 98.08% recall, and a 98.12% F1 score for multi-class classification |
| Abhijeet Patil et al. [28] | BreaKHis, BACH | Attention-based Multiple Instance Learning (A-MIL) | Lower classification accuracy compared to state-of-the-art methods; computational complexity from multiple instance approach; tested on limited magnification levels only. | Achieved classification accuracy of 86.56% at 200x magnification and effective localization of malignant regions |
| Wei Wang et al. [29] | BreaKHis, BUSI | Semi-supervised Vision Transformer (ViT) with Adaptive Token Sampling (ATS) | Needs labeled data for semi-supervised learning; adaptive token sampling may miss subtle features; only tested on two datasets; potential overfitting to common breast cancer patterns. | Achieved 98.12% accuracy, 98.17% precision, 98.65% recall, and 98.41% F1-score on BreaKHis |

## Chapter 3 Methodology

### 3.1 Approach

Building upon the challenges and research gaps highlighted in the background review, this chapter presents the comprehensive methodology adopted for developing and validating the proposed hybrid deep learning framework. The approach integrates advanced Convolutional Neural Networks (CNNs) and Vision Transformer (ViT) models, supported by robust preprocessing pipelines and performance evaluation protocols. The structure of this chapter is organized to guide the reader through the logical and technical progression of the project, ensuring clarity and reproducibility of results.

### 3.2 Dataset

#### 3.2.1 Dataset1—Breast Cancer Histopathological Database (BreakHis)

The Breast Cancer Histopathological Database (BreakHis) [15] was employed in this study, comprising 7,909 microscopic images of breast tumour tissue samples obtained through biopsy procedures. These images were captured at various magnification factors (40X, 100X, 200X, and 400X), with this study specifically focusing on the 100X magnification level, which includes 1,995 benign and 2,081 malignant samples. The dataset is organized into two main classes: benign and malignant tumours, providing a comprehensive foundation for binary classification of breast cancer histopathological images.

For the dataset organization, let $D_{total}$ represent the complete dataset:

$$\mathrm{D_{total}} = \mathrm{D_{train}} \cup \mathrm{D_{val}} \cup \mathrm{D_{test}} \tag{1}$$

where $|D_{train}|:|D_{val}|:|D_{test}| = 0.7:0.1:0.2$

To ensure robust model evaluation, this project implemented a custom data split strategy that differs from the original dataset organization. Rather than using the predefined train/test split, this project adopted a more comprehensive approach with a 70-30 split for initial separation, followed by further dividing the 30% portion into validation and test sets(10-20). For each subset $D_i$, this project maintains class balance through:

$$\frac{\left|D_i^{benign}\right|}{\left|D_i^{benign}\right| + \left|D_i^{malignant}\right|} \approx \frac{\left|D_{total}^{benign}\right|}{|D_{total}|} \tag{2}$$

To address class imbalance issues in the training set, this project applied random oversampling techniques. The oversampling process can be expressed as:

$$\left|D_{train}^{benign}\right| = \left|D_{train}^{malignant}\right| = max\left(\left|D_{train}^{benign}\right|, \left|D_{train}^{malignant}\right|\right) \quad (3)$$

This ensures a balanced distribution of classes for model training while maintaining the integrity of the original samples.

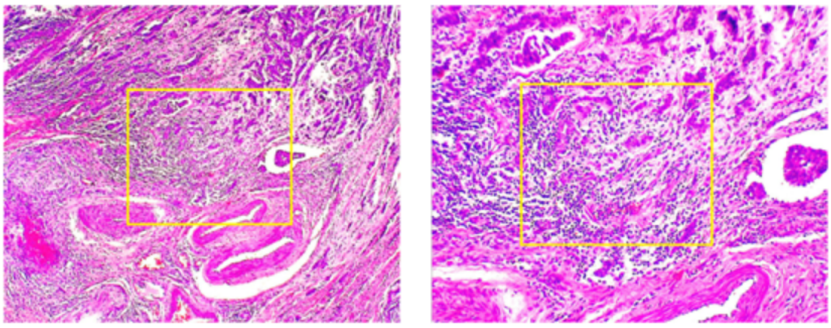

Figure 5 Sample of BreaKHis Dataset

### 3.2.2 Dataset2—ICIAR2018_BACH_Challenge dataset

To improve the model's generalization across different histopathological conditions, the ICIAR2018_BACH_Challenge dataset (hereafter referred to as BACH)[30] was also incorporated into this study. The original dataset consists of 400 high-resolution microscopy images, each measuring 2048×1536 pixels, divided equally across four categories: Normal, Benign, In Situ Carcinoma, and Invasive Carcinoma as illustrated in Figure 6.

For consistency with the binary classification task defined in Section 3.2.1, this project excluded the *Normal* category and redefined the remaining classes as follows:

- **Benign (Class 0):** All samples originally labelled *Benign*
- **Malignant (Class 1):** All samples originally labelled *In Situ Carcinoma* or *Invasive Carcinoma*

Let $D_{\text{bach}}$ represent the refined BACH dataset, with total samples $N = 300$, where:

$$D_{\text{bach}} = D_{\text{benign}} \cup D_{\text{malignant}}, \left|D_{\text{benign}}\right| = 100, \left|D_{\text{malignant}}\right| = 200 \quad (4)$$

To prepare the dataset for training, this project employed an 80–20 stratified split to form training and testing subsets, denoted as $D_{\text{train}}^{\text{bach}}$ and $D_{\text{test}}^{\text{bach}}$ respectively, satisfying:

$$\left|D_{\text{train}}^{\text{bach}}\right| : \left|D_{\text{test}}^{\text{bach}}\right| = 0.8 : 0.2 \tag{5}$$

This results in:

$$\left|D_{\text{train}}^{\text{bach}}\right| = 240, \left|D_{\text{test}}^{\text{bach}}\right| = 60 \tag{6}$$

To address the class imbalance inherent in the training set, this project applied random oversampling so that:

$$\left|D_{\text{train}}^{\text{benign}}\right| = \left|D_{\text{train}}^{\text{malignant}}\right| = \max\left(\left|D_{\text{train}}^{\text{benign}}\right|, \left|D_{\text{train}}^{\text{malignant}}\right|\right) = 160 \tag{7}$$

Resulting in a balanced training set $D_{\text{train}}^{\text{balanced}}$ of 320 samples.

Each image was resized to $160 \times 160$ pixels and normalized to the range $[0,1]$ to match the BreakHis preprocessing pipeline and facilitate joint training. Class labels were encoded as 0 for benign and 1 for malignant cases:

$$y = \begin{cases} 0, & if\ class = Benign \\ 1, & if class \in InSitu, Invasive \end{cases} \tag{8}$$

This harmonized preprocessing ensured compatibility across datasets and minimized domain shift, allowing the BACH data to serve as an effective supplement to the BreakHis dataset during model training and evaluation.

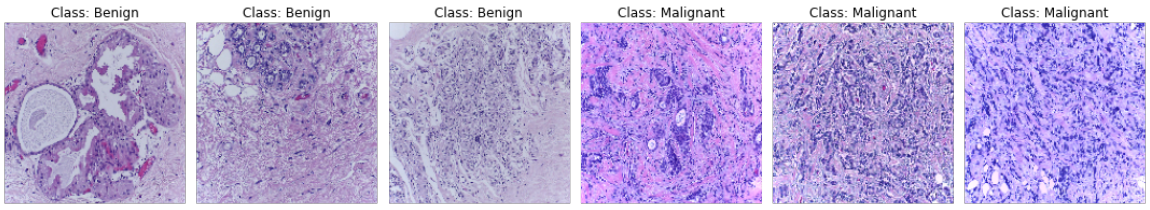


Figure 6 Sample of BACH Dataset

### 3.2.3 Data Preprocessing

The data preprocessing pipeline was carefully designed to optimize the dataset for deep learning model training. Each image underwent several crucial preprocessing steps.

#### i. Data Normalization

First, the images were resized to a uniform dimension of 160×160 pixels, chosen to balance computational efficiency with the preservation of important histopathological features. For an input image $I$, the normalization process can be expressed as:

$$I_{normalized} = \frac{I - min(I)}{max(I) - min(I)} \quad (9)$$

### ii. Data Augmentation

To enhance model robustness and prevent overfitting, this project implemented comprehensive data augmentation techniques. The augmentation transformations can be represented as a composition of functions:

$$I_{augmented} = T_n \circ T_{n-1} \circ \ldots \circ T_1(I) \quad (10)$$

where each transformation $T_i$ belongs to the following set of possible augmentations:

- Rotation: $T_{rot}(I, \theta)$ where $\theta \in [-90°, 90°]$
- Zoom: $T_{zoom}(I, z)$ where $z \in [0.8, 1.2]$
- Translation: $T_{trans}(I, \triangle x, \triangle y)$ where $\triangle x, \triangle y \in [-0.2w, 0.2w]$, and $w$ is the image width
- Flip: $T_{flip}(I) \in \{I, I_{horizontal}, I_{vertical}\}$

The preprocessing pipeline was integrated into the training process using TensorFlow's data pipeline, allowing for efficient batch processing and real-time augmentation during training. This approach not only ensures efficient memory usage but also provides a continuous stream of varied training examples, contributing to better model generalization. Each processed ima $I \in R^{160 \times 160 \times 3}$ is properly normalized to $[0,1]^{160 \times 160 \times 3}$ and augmented when necessary, while maintaining the integrity of the evaluation process through consistent normalization across all sets.

To provide visual insight into the preprocessing pipeline, representative samples from both the BreakHis and BACH datasets are presented below. Each image has been resized to $160 \times 160$ pixels, normalized to $[0,1]$, and subjected to random augmentation. The selected samples include one benign and one malignant case from each dataset.

**BreakHis Dataset (100× Magnification):**

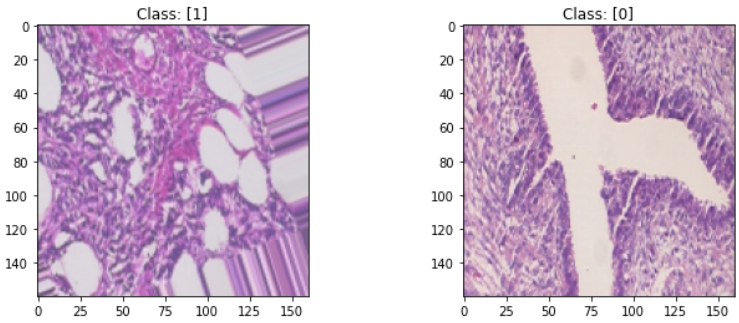


Figure 7 Preprocessed Samples from BreakHis Dataset(100x)

**BACH Dataset:**

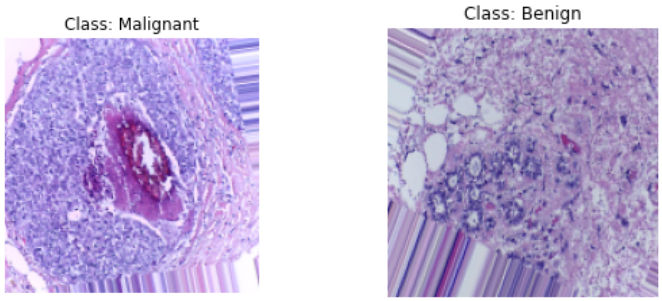


Figure 8 Preprocessed Samples from the BACH Dataset

This figure illustrates one benign and one malignant histopathological image from each dataset following the complete preprocessing pipeline. All images have been resized to 160×160 pixels, normalized to a pixel intensity range of [0, 1], and may include data augmentation transformations such as rotation, zoom, translation, or flipping. The BreakHis samples (top row) are captured at 100× magnification, while the BACH samples (bottom row) originate from the ICIAR2018 Grand Challenge dataset. These visual examples highlight the morphological differences between benign and malignant tissues as interpreted by the model.

## 3.3 Model Design and Evaluation

### 3.3.1 EfficientNet-based Convolutional Neural Network

This research implements an EfficientNetV2-B0 [31] based architecture for processing histopathological images, leveraging its efficient compound scaling and advanced convolution operations. The model architecture enhances feature extraction by capturing spatial information at multiple scales, allowing for the detection of both fine-grained cellular structures and coarse tissue organization patterns.

At the core of the architecture, the convolutional operation processes the input image through learnable filters. For an input image region $I$ and kernel $K$, the basic convolution operation is defined as:

$$f(i,j) = \sum_m \sum_n I(i+m, j+n) \cdot K(m,n) \tag{11}$$

where $f(i,j)$ represents the resulting feature map value at position $(i,j)$. The EfficientNetV2 architecture employs several advanced convolution variants, including:

1. MBConv (Mobile Inverted Bottleneck Convolution):

$$MBConv(X) = PWC\left(DWC\big(PWC(X)\big)\right) \tag{12}$$

where $PWC$ represents pointwise convolution and $DWC$ represents depthwise convolution.

2. Fused-MBConv:

$$FusedMBConv(X) = PWC\big(Conv(X)\big) \tag{13}$$

The network architecture consists of multiple stages, with each stage operating at different spatial resolutions. The feature extraction process can be represented as:

$$F_l = H_l(F_{l-1}) \tag{14}$$

where $F_l$ represents the feature maps at layer $l$, and $H_l$ is the composite transformation at that layer.

The bottleneck structure of the network is designed with:

$$X_{out} = X_{in} + \phi\left(BN\left(Conv\left(BN\big(Conv(X_{in})\big)\right)\right)\right) \tag{15}$$

where $BN$ represents batch normalization, and $\phi$ is the activation function (Swish/SiLU):

$$\phi(x) = x \cdot sigmoid(x) \tag{16}$$

For optimization, the model employs the Adam optimizer with an inverse time decay learning rate schedule:

$$\alpha_t = \alpha_0 \cdot \frac{1}{1 + \beta t} \tag{17}$$

where $\alpha_0 = 0.001$ is the initial learning rate, $\beta$ is the decay rate, and $t$ is the current training step.

The loss function utilizes sparse categorical cross-entropy:

$$L = -\sum_{c=1}^{M} y_c \log(\hat{y_c}) \tag{18}$$

where $y_c$ represents the true label, $\hat{y_c}$ is the predicted probability for class $c$, and $M = 2$ for the binary classification task.

The architecture incorporates several regularization techniques to prevent overfitting:

- Dropout with probability $p = 0.7$
- L2 regularization with weight decay $\lambda = 0.0001$
- Batch normalization with momentum $\mu = 0.99$

The feature extraction pathway processes input images of size 160×160×3 through sequential blocks of convolutions and pooling operations. Each block progressively reduces spatial dimensions while increasing the feature channel depth, following the principle:

$$C_{out} = \alpha \cdot C_{in} \tag{19}$$

$$H_{out} = \frac{H_{in}}{\rho} \tag{20}$$

$$W_{out} = \frac{W_{in}}{\rho} \tag{21}$$

where $\alpha$ represents the channel multiplier and $\rho$ represents the reduction factor for spatial dimensions.

The final architecture produces a rich feature representation that captures multi-scale patterns in histopathological images. This hierarchical feature extraction proves particularly effective for identifying the complex structural patterns characteristic of benign and malignant breast tissue samples. The network's efficiency in both computational resources and parameter utilization makes it well-suited for medical image analysis tasks where both accuracy and processing speed are crucial considerations.

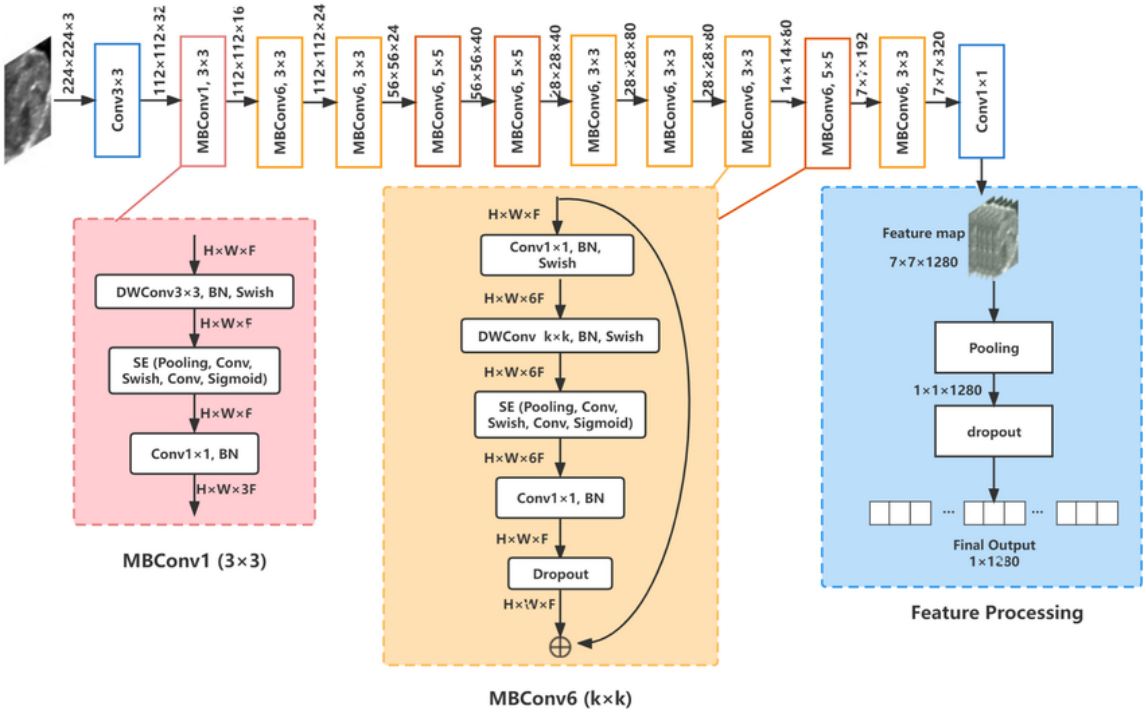


Figure 9 Architecture of EfficientNetV2-B0 Backbone

Figure 9 illustrates the overall structure of the EfficientNetV2-B0 model, including stacked MBConv and Fused-MBConv blocks at varying spatial resolutions. Each block incorporates depthwise separable convolutions, squeeze-and-excitation modules, batch normalization, and Swish activation. The feature extraction culminates in global average pooling and projection to the final embedding, suitable for binary classification tasks on histopathological images.

### 3.3.2 Vision Transformer (ViT)

The Vision Transformer (ViT) architecture [32] implemented in this research aims to capture long-range dependencies and provide a global contextual understanding of histopathological images. Unlike CNNs, which primarily focus on local features, ViT

processes images by first dividing them into fixed-size patches. For an input image $I \in R^{H\times W\times C}$, the patch tokenization process creates a sequence of $N$ patches:

$$P = \{p_1, p_2, \dots, p_N\} \text{ where } N = \frac{HW}{P^2} \tag{22}$$

where each patch $p_i \in R^{P\times P\times C}$, with $P = 16$ being the patch size in this implementation.

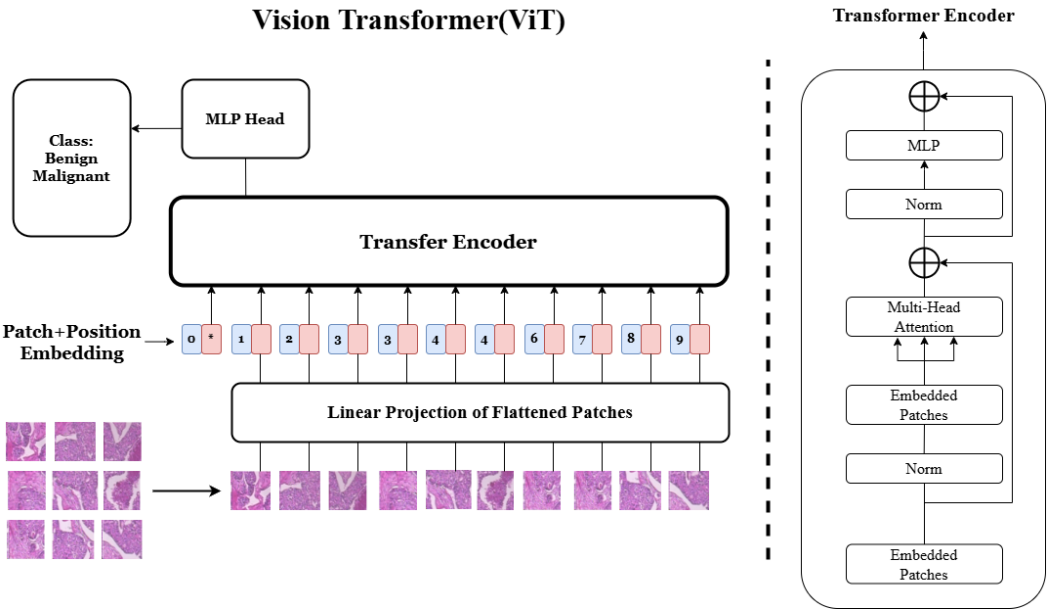


Figure 10 Architecture of the Vision Transformer(ViT) Model

The standard Vision Transformer (ViT) architecture forms the foundation of the transformer-based modeling used in this research. Following patch tokenization and position embedding, the input sequence is processed through multiple layers of multi-head self-attention and feed-forward sub-networks. Each patch serves as a token, and a learnable classification token is prepended to the sequence. The model's attention mechanism enables it to capture long-range dependencies across the entire image, a capability particularly valuable in medical image analysis where pathological structures may span distant regions. Figure 10 above provides a high-level illustration of the ViT model employed in this study, highlighting the token embedding process, transformer encoder stack, and the MLP classification head used for binary prediction.

This research enhances the standard ViT architecture with Shifted Patch Tokenization (SPT), which creates multiple views of the input image through spatial shifting:

$$I_{shifted} = \{I, I_{left-up}, I_{left-down}, I_{right-up}, I_{right-down}\} \quad (23)$$

Each patch is then flattened and linearly projected to create token embeddings. Position embeddings are added to maintain spatial information:

$$z_0 = \left[x_{class}; x_p^1 E; x_p^2 E; \ldots; x_p^N E\right] + E_{pos} \quad (24)$$

where $E$ is the patch embedding matrix and $E_{pos}$ represents learnable position embeddings.

The core of the ViT architecture is the self-attention mechanism, implemented through Multi-Head Self-Attention (MSA) layers. For each attention head, the attention operation is computed as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (25)$$

where $Q$, $K$, and $V$ are query, key, and value matrices derived from the input embeddings, and $d_k = 64$ is the dimensionality of the key vectors. The multi-head attention combines outputs from multiple attention heads:

$$MSA(X) = Concat(head_1, \ldots, head_h)W^O \quad (26)$$

$$head_i = Attention\left(XW_i^Q, XW_i^K, XW_i^V\right) \quad (27)$$

The transformer encoder consists of multiple layers ($L = 4$), each containing MSA and feed-forward network (FFN) blocks with layer normalization (LN):

$$x' = MSA\big(LN(x)\big) + x \quad (28)$$

$$x'' = FFN\big(LN(x')\big) + x' \quad (29)$$

The feed-forward network applies two linear transformations with a GELU activation:

$$FFN(x) = W_2\sigma(W_1 x + b_1) + b_2 \quad (30)$$

where σ is the GELU activation function.

### 3.3.3 Shifted Patch Tokenization (SPT)

To improve the ViT model's capacity for fine-grained spatial understanding, this study integrates a Shifted Patch Tokenization (SPT) mechanism [33] as a preprocessing enhancement. Instead of relying solely on a single static view of the input image, SPT introduces spatial transformations that generate four additional versions of the image through directional shifts. These shifted variants are then concatenated at the channel level and partitioned into fixed-size patches, producing enriched patch embeddings that encode both global and localized visual cues. This multi-perspective approach increases the diversity of patterns available to the attention mechanism, particularly benefitting tasks such as histopathological cancer detection where subtle changes in texture or structure are diagnostically significant. Figure 11 illustrates the full SPT pipeline, from spatial transformations through linear projection to the resulting visual tokens.

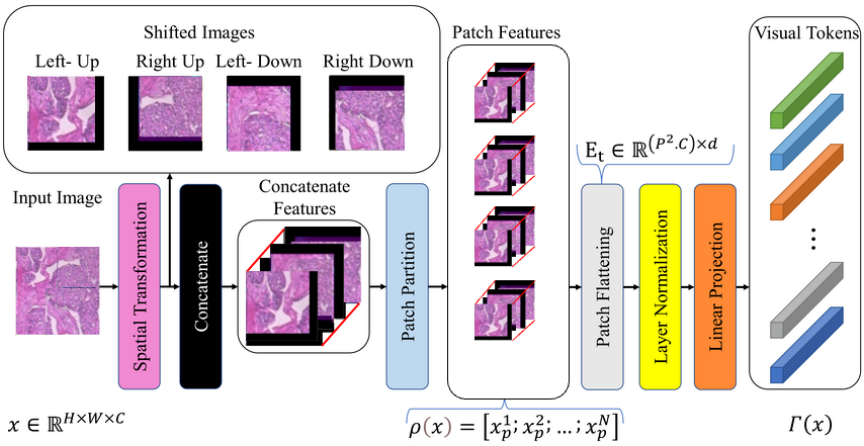


Figure 11 Shifted Patch Tokenization(SPT) Framework

SPT patch processing pseudocode:

**Algorithm** Shifted Patch Tokenization (SPT)

```
Initialize empty list tokens
for direction in ['original', 'up', 'down', 'left', 'right'] do
    shifted_img ← tf.roll(image, shift=Δ, axis=direction)
    patches ← extract_patches(shifted_img)
    embedded ← Dense(64)(patches)
    tokens.append(embedded)
end for
final_tokens ← Concatenate(tokens, axis=channels)
```

### 3.3.4 Locality Self-Attention (LSA)

The Learned-Scale Attention (LSA) mechanism [33] was implemented to enhance the model's flexibility in distributing attention across tokens when analyzing complex histopathological images. In standard self-attention, the attention logits are scaled by the fixed factor $\sqrt{d_k}$ to stabilize gradients. However, a static scaling may not be optimal across diverse data distributions.

In the LSA approach, this static scaling factor is replaced by a learnable temperature parameter $\tau$, initialized to $\sqrt{d_k}$ but updated during training. This modification enables the model to dynamically adjust the sharpness of the attention distribution, allowing it to better emphasize important tissue regions while suppressing less relevant features. The overall structure of the attention computation remains consistent with the Vision Transformer framework described earlier, with the key difference being the learnable scaling of logits before the softmax operation.

Practically, the LSA module was implemented by subclassing the standard Multi-Head Attention layer in TensorFlow/Keras. The temperature parameter $\tau$ is trained jointly with other model parameters through backpropagation, allowing the model to automatically adapt its attention behavior based on the characteristics of histopathological data.

Figure 12 illustrates the workflow of the Learned-Scale Attention mechanism. By scaling attention logits with a learnable temperature, the model can produce smooth, weakly sharp, or strongly sharp attention distributions depending on the needs of the input features. This flexibility significantly improves the model's ability to capture subtle but critical patterns for early cancer detection.

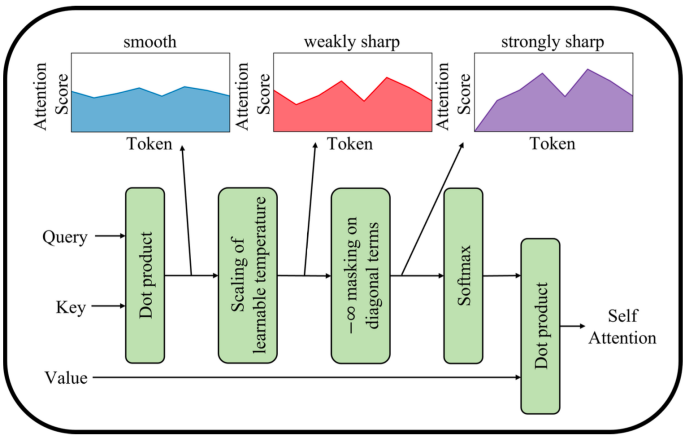


Figure 12 Workflow of the Learned-Scale Attention (LSA) mechanism

SPT patch processing pseudocode:

**Algorithm** Learned-Scale Attention (LSA)

1. Initialize learnable temperature parameter $\tau \leftarrow \sqrt{d_k}$
2. Obtain query $Q$, key $K$, and value $V$ from input embeddings
3. Compute attention logits: $scores \leftarrow QK^T$
4. Scale attention logits: $scaled\_scores \leftarrow scores \,/\, \tau$
5. Apply masking on diagonal terms if necessary
6. Normalize with softmax: $attention\_weights \leftarrow softmax(scaled\_scores)$
7. Compute output: $output \leftarrow attention\_weights \times V$
8. Return output

### 3.3.5 Model Overview

To provide a clearer understanding of the proposed system architecture, this section outlines the overall structure of the hybrid model integrating convolutional and transformer-based components. The design leverages the local feature extraction capabilities of EfficientNetV2 in the CNN branch and the global contextual modelling strength of Vision Transformers (ViT) in the transformer branch. These two parallel pathways are fused at the feature level and followed by fully connected layers to produce the final classification. The full architecture is illustrated in Figure 13.

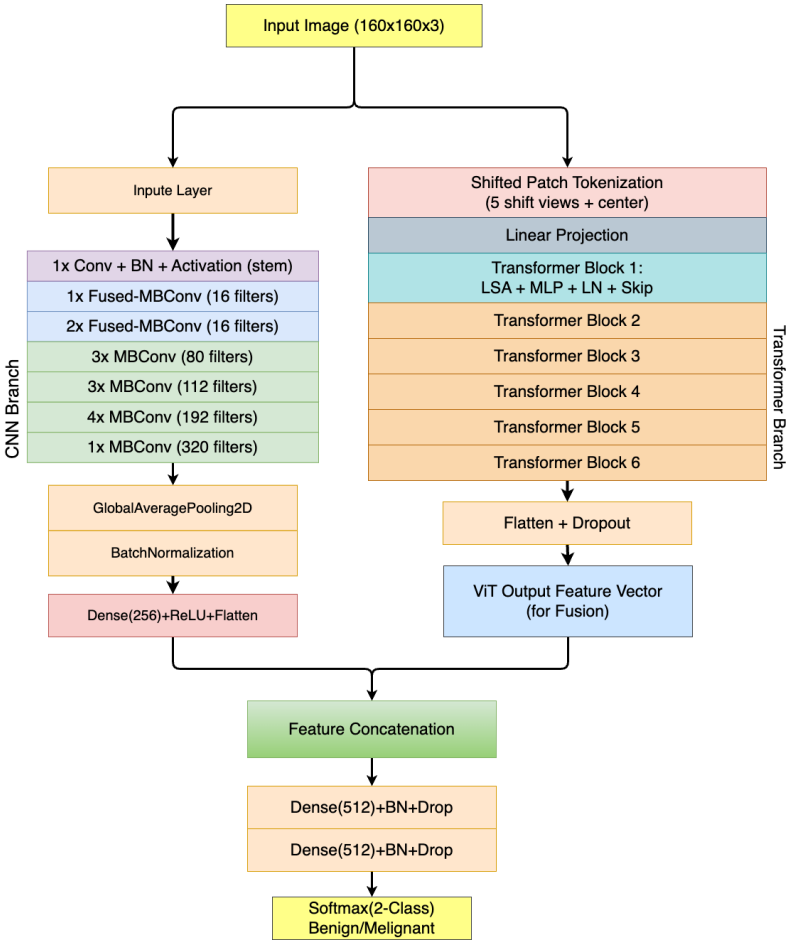


Figure 13 Overall Architecture of Hybrid CNN-ViT Model

### 3.3.6 Model Optimization and Classification Strategy

The model is optimized using the Adam optimizer with an initial learning rate of $\alpha_0 = 0.001$ and a weight decay of $\lambda = 0.0001$. The learning rate follows an inverse time decay schedule:

$$\alpha_t = \alpha_0 \cdot \frac{1}{1 + \beta t} \tag{31}$$

where $\beta$ is the decay rate.

To prevent overfitting, several regularization techniques are implemented:

- Dropout with rate $p = 0.7$
- Layer normalization with $\epsilon = 1e - 6$
- Weight decay regularization

The final classification head consists of:

1. Layer normalization
2. Global average pooling over patch embeddings
3. MLP layers with dimensions [512, 128]
4. Final classification layer

### 3.3.7 Evaluation Metrics

To comprehensively evaluate the performance of the proposed classification models, a multi-metric evaluation framework was implemented. Given the medical context of the task, detecting malignant breast cancer tissue in histopathological images, it is essential to move beyond simple accuracy and consider metrics that provide deeper insights into class-wise behavior, especially for identifying false negatives.

The primary evaluation metrics used in this study include accuracy, precision, recall, F1-score, and AUC-ROC, defined mathematically as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{32}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{33}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{34}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{35}$$

Here, TP (true positives) and TN (true negatives) represent the number of correctly classified malignant and benign cases, respectively, while FP (false positives) and FN (false negatives) denote misclassifications. These four values form the basis of all evaluation metrics in binary classification tasks.

Each metric provides a different perspective on model performance:

- Accuracy indicates the overall correctness of the model but can be misleading in imbalanced datasets where one class dominates.

- Precision quantifies the proportion of correctly identified positive cases among all predicted positives. High precision means fewer false alarms which is critical in reducing unnecessary follow-up procedures in medical settings.
- Recall reflects the model's ability to capture actual positive instances. In the context of cancer detection, high recall is crucial to minimize missed malignancy diagnoses.
- F1-Score serves as the harmonic mean of precision and recall. It provides a single score that balances both concerns, especially useful when precision and recall are in tension.
- AUC-ROC (Area Under the Receiver Operating Characteristic Curve) evaluates the model's overall ability to discriminate between classes across various threshold settings. A higher AUC indicates a stronger capability to distinguish malignant from benign samples regardless of the decision boundary.

To monitor model learning dynamics and convergence during training, the binary cross-entropy loss function is employed:

$$L(\theta) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \quad (36)$$

Here, $\theta$ denotes the trainable model parameters, $y_i \in \{0,1\}$ is the true binary label, and $\hat{y}_i \in [0,1]$ is the predicted probability for class 1. Tracking the loss over time helps assess whether the model is converging stably or exhibiting symptoms of overfitting or underfitting.

All metrics are computed on an independent hold-out test set, which was separated prior to training and never exposed to the model during optimization or validation. This ensures that the reported performance reflects the true generalization capability of the model.

### 3.3.8 Model Explainability

Model explainability is a crucial aspect of machine learning, especially when dealing with high-stakes applications such as medical diagnosis. In this project, this project adopt Grad-CAM (Gradient-weighted Class Activation Mapping) as a key technique to enhance the interpretability of the hybrid deep learning model for breast cancer detection. Grad-CAM provides a visual representation of the regions in an image that are most influential in the model's decision-making process, thereby allowing for better insight into the model's behavior.

Grad-CAM works by utilizing the gradients of the target class with respect to the convolutional layer's output. These gradients are then used to generate a heatmap that highlights important regions of the input image. This process helps to identify which parts of the image the model is focusing on when making a classification, offering valuable explanations that are especially critical in medical image analysis, where the trust and transparency of the model's decisions are paramount.

Mathematically, the Grad-CAM heatmap can be expressed as follows:

$$\text{Heatmap} = \text{ReLU}\left(\sum_{k} \alpha_k A^k\right) \tag{37}$$

where:

- $A^k$ represents the activation map from the convolutional layer for a specific filter $k$,
- $\alpha_k$ denotes the weight associated with each filter, calculated as the average of the gradients with respect to the class score, i.e., $\alpha_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial y_c}{\partial A_{ij}^k}$,
- $y_c$ is the score for class c,
- $Z$ is a normalization factor to ensure that the weights are appropriately scaled.

Grad-CAM not only enhances the interpretability of the model by providing visual insights into the decision-making process but also helps in validating the model's focus areas. In the context of breast cancer detection, for example, the heatmap can highlight regions of malignant tissue that the model is focusing on, offering confidence to clinicians about the model's reasoning. This is particularly beneficial for cases where subtle features, such as small malignant regions or abnormal tissue textures, play a crucial role in the detection process.

Moreover, Grad-CAM can be extended to analyze the importance of specific regions in the image by masking parts of the image and observing the changes in model performance. By selectively masking regions identified as important, can evaluate the model's sensitivity to different areas and determine whether it is relying on biologically relevant features for classification. This ability to visualize the impact of different regions allows for further refinement of the model, enhancing its reliability and robustness in clinical applications.

In this study, Grad-CAM is integrated into the model's evaluation pipeline to not only improve transparency but also to serve as a diagnostic tool for model validation. The effectiveness of Grad-CAM in highlighting meaningful features provides additional validation that the model is capable of learning features relevant to early cancer detection, rather than relying on spurious correlations or irrelevant patterns.

By providing these visual explanations, Grad-CAM contributes to a more interpretable model that can be confidently used in sensitive fields like healthcare, where understanding the reasoning behind automated predictions is as important as the predictions themselves.

The heatmap is then upscaled to match the original input image dimensions and applied to the image to create a visualization that emphasizes the regions that most strongly contribute to the model's prediction.

### 3.4 Environmental Setup and Technology

The project will be developed using a combination of local and remote environments to facilitate both model development and large-scale training tasks. Locally, the setup consists of a Windows 10 operating system, utilizing Visual Studio Code for development, with TensorFlow 2.13 and common Python libraries like Keras, NumPy, Pandas, and Matplotlib for deep learning operations and data handling. The local hardware includes a NVIDIA RTX 3060 GPU and an AMD Ryzen 7 5800H CPU, suitable for small-scale experiments.

For more computationally intensive tasks, a remote Linux server (Ubuntu) environment will be used. The remote server provides flexibility with Jupyter Lab for development, and its hardware is configurable to support multiple GPUs as needed for faster and parallelized processing during deep learning model training.

Table 2 Configure hardware and software resources

| **Resource** | **Local Environment** | **Remote Server** |
|---|---|---|
| Operating System | Windows 10 | Linux Ubuntu |
| Software | Visual Studio Code | Jupyter Lab, Linux Terminal |
| Python Libraries | TensorFlow 2.13, Keras, NumPy, Pandas, Matplotlib, etc. | Customizable: TensorFlow, Keras, NumPy, etc. |
| GPU | NVIDIA RTX 3060 | Configurable to support multiple NVIDIA RTX 4090 GPUs |
| CPU | AMD Ryzen 7 5800H | Configurable |

| | | |
|---|---|---|
| Others | PyCharm IDE for local development | Jupyter Lab for remote experimentation |

### 3.5 Project Version Management

To ensure efficient collaboration, reproducibility, and traceability of development progress, version control and project source management were conducted using Git and hosted on GitHub**.** The repository maintained all essential components of the project, including model architecture definitions, training scripts, preprocessing pipelines, evaluation utilities, and the GUI deployment interface.

Additionally, all relevant project documentation including the original project proposal and the mid-term project progress report was uploaded to the repository. This ensures full transparency in the project's evolution and allows future researchers to reference the initial design rationale and mid-course adjustments.

The GitHub platform was used throughout the project lifecycle to:

- Track incremental changes through commits with meaningful messages;
- Document experiments and parameter tuning trials using Jupyter notebooks;
- Manage model weights, configuration files, and datasets via .gitignore and external cloud links to avoid bloating the repository;
- Provide a centralized and accessible version of the source code for external review or deployment.

The project repository is publicly accessible at the following link:

https://github.com/CarsonLLuo/FinalProject

In addition to GitHub, all project resources—from the initial project proposal onward—have been systematically backed up and archived using Baidu NetDisk for redundancy and long-term accessibility. This includes intermediate files, supplementary materials, and internal documentation. The Baidu NetDisk archive is accessible via the following link:

https://pan.baidu.com/s/1i9IYGIKLgzlypq0gPsKo0g?pwd=ua9y

## Chapter 4 Implementation and Results

This chapter presents the detailed implementation of the proposed model, following the methodology outlined in Chapter 3. The model leverages a hybrid architecture that

combines EfficientNetV2-B0 for feature extraction with the Vision Transformer (ViT), enhanced by Shifted Patch Tokenization (SPT) to capture both fine-grained and global features of histopathological images. The implementation process encompasses data preprocessing, model architecture design, training setup, and performance evaluation. The following section provides an in-depth overview of each implementation step, beginning with the data processing pipeline that prepares the input for model training.

## 4.1 Model Implementation

This section presents the implementation details of the proposed hybrid deep learning model, focusing on the engineering and programming aspects of each module. While the architectural rationale has already been discussed in Chapter 3, here this project describes how each module was constructed, trained, and integrated using TensorFlow and Keras.

### 4.1.1 EfficientNetV2 Backbone

The EfficientNetV2-B0 architecture was adopted for its balance between performance and computational efficiency, particularly on medical image classification tasks. In this implementation, the model was loaded without the final classification layer to allow custom adaptation. Input histopathological images were uniformly resized to 160×160×3 and then fed into the backbone network.

After passing through the backbone, the output feature maps were compressed and refined through several additional layers. A global average pooling layer was applied first to reduce spatial dimensions, followed by batch normalization to stabilize training. A fully connected layer with ReLU activation was then used to project the features into a lower-dimensional space. Finally, the output was flattened into a one-dimensional vector to serve as the input for the subsequent fusion stage with the Vision Transformer branch.

### 4.1.2 Vision Transformer Backbone

The Vision Transformer (ViT) module was custom-built to capture long-range dependencies and global context within histopathological images. Each input image was divided into fixed-size non-overlapping patches of 16×16 pixels. These patches were individually flattened and linearly projected into a 64-dimensional embedding space. To retain spatial structure, learnable positional encodings were added to the patch embeddings.

The core ViT architecture consisted of several stacked encoder blocks. Each block integrated multi-head self-attention with four attention heads and a key dimension of 64, followed by a feed-forward multilayer perceptron consisting of two dense layers with hidden dimensions of 128 and 64, respectively, activated by GELU. Layer normalization and residual connections were incorporated to facilitate stable optimization and efficient gradient flow.

### 4.1.3 Shifted Patch Tokenization (SPT)

To enrich the spatial representation of input images, a Shifted Patch Tokenization (SPT) mechanism was introduced. This technique generates multiple shifted versions of the original image by applying pixel-level rolling operations along different directions, including upward, downward, leftward, and rightward shifts.

Each shifted image was subsequently divided into non-overlapping patches, which were flattened and embedded into a common 64-dimensional space. The resulting token sequences from all shifted views were then concatenated along the channel dimension, creating a more diverse and information-rich representation. This augmentation enhances the model's ability to capture fine-grained variations in tissue structures, leading to improved performance in histopathological image classification.

### 4.1.4 Learned-Scale Attention (LSA)

To enhance the flexibility of the attention mechanism within the Vision Transformer, a Learned-Scale Attention (LSA) strategy was implemented. In conventional Transformer models, attention scores are normalized using a static scaling factor based on the embedding dimension. However, this approach may constrain the model's ability to adaptively regulate the sharpness of attention distributions, which is crucial for distinguishing subtle patterns in histopathological images.

In this implementation, the scaling factor was redefined as a trainable parameter, initialized to a value consistent with standard normalization practices. During model training, this parameter was optimized alongside other network weights, allowing the model to learn the most effective scaling strategy based on data characteristics. The learned scaling parameter governs the relative focus of attention heads, enabling either sharper or more diffuse attention distributions depending on the demands of the feature space.

The LSA mechanism was integrated seamlessly within the multi-head self-attention blocks of the Transformer encoder. Without altering the overall attention computation flow, it introduced an additional degree of freedom that allowed attention scores to self-adjust dynamically over the course of training. By learning an optimal balance between focus and dispersion, the LSA-enhanced Transformer demonstrated improved representational capacity, leading to better discrimination between benign and malignant histopathological samples.

### 4.1.5 Hybrid Feature Fusion

To effectively leverage both local and global information, the feature vectors produced by the EfficientNetV2 backbone and the Vision Transformer module were concatenated to form a unified representation. This fused vector encapsulates both fine-grained cellular characteristics and broader contextual patterns across the histopathological images.

The concatenated features were subsequently passed through a sequence of fully connected layers to enable joint learning. The first dense layer comprised 512 units with ReLU activation, followed by a dropout layer to reduce overfitting. This was followed by another dense layer with 128 units and ReLU activation, along with a second dropout layer. Finally, a softmax-activated output layer with two units was used to perform binary classification between benign and malignant samples.

### 4.1.6 Training Configuration

The hybrid model was trained using the Adam optimizer with an inverse time decay learning rate schedule. The total training duration spanned 200 epochs, with early stopping applied to halt training after 10 epochs of no improvement in validation loss.

All training was executed on a Linux server with an NVIDIA RTX 4090 GPU.

**Training Hyperparameters**

Table 3 Hyperparameters Setting

| Hyperparameter | Value |
|---|---|
| Optimizer | Adam |
| Initial Learning Rate | 0.001 |
| Learning Rate Decay | 1e-3 |
| Batch Size | 32 |
| Epochs | 200 |

| | |
|---|---|
| Dropout Rate | 0.2 |
| L2 Regularization | 1e-4 |
| MLP Dimensions (ViT) | [128, 64] |
| Early Stopping Patience | 10 |

## 4.2 Result Analysis

In this section, this project present a detailed analysis of the model's performance across different evaluation metrics, focusing on its ability to classify histopathological images into benign and malignant categories. This project will begin by discussing the performance of the hybrid model on both the BreakHis and BACH datasets, followed by a comparison of key metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. These metrics provide a comprehensive understanding of how well the model distinguishes between benign and malignant tissues.

### 4.2.1 Performance on BreakHis Dataset

The model's performance on the BreakHis dataset is evaluated across multiple key metrics. The evaluation begins with an analysis of the accuracy and loss curves, followed by a detailed breakdown of the confusion matrix. This project discuss the precision, recall, F1-score, and visualize these metrics with a corresponding F1 vs. Recall curve. Finally, this project assess the ROC curve and AUC, providing a comprehensive view of the model's performance in classifying benign and malignant tissue samples.

#### 4.2.1.1 Accuracy and Loss Curves

The training and validation curves are illustrated in Figure 14. The left plot presents the accuracy trends, where training accuracy increased steadily and reached 0.9317, while validation accuracy initially fluctuated before stabilizing at 0.9095, indicating satisfactory generalization after early convergence.

The right plot depicts the loss curves. Training loss exhibited a consistent downward trend, converging at 0.8145. Validation loss initially rose slightly but eventually decreased to 0.8934, suggesting effective learning with mild overfitting risk in the early epochs.

Overall, the model demonstrated stable convergence, with high accuracy and minimal overfitting, indicating strong generalization performance on the validation set.

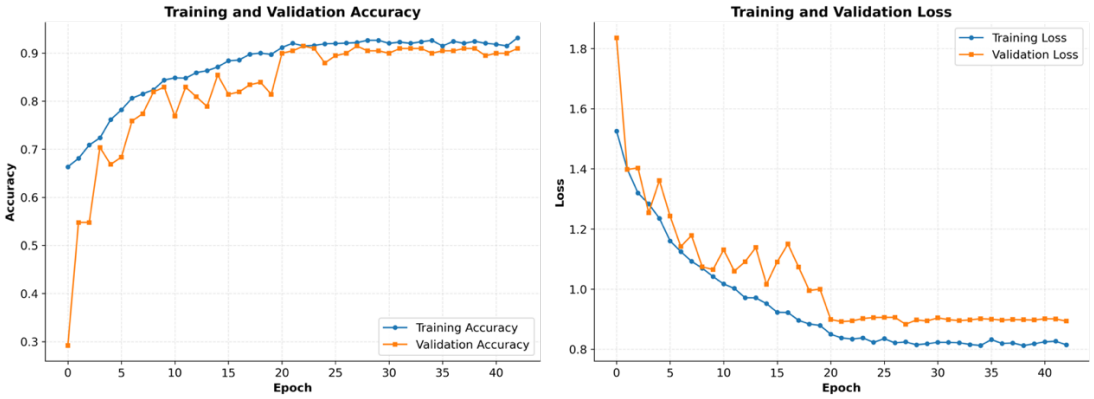


Figure 14 Accuracy (left) and Loss (right) over Epochs for BreakHis Dataset

#### 4.2.1.2 Confusion Matrix

The confusion matrix in Figure 15 summarizes the model's performance on the BreakHis dataset. The model correctly classified 107 true negatives (TN) and 265 true positives (TP), corresponding to benign and malignant samples respectively. Misclassifications included 18 false positives (FP), where benign samples were predicted as malignant, and 15 false negatives (FN), where malignant samples were predicted as benign.

These results indicate strong classification performance, with the model achieving high true positive and true negative counts across both classes. The relatively low number of false positives and false negatives suggests that the model generalizes well and maintains a balanced sensitivity and specificity, which is critical for medical diagnosis tasks.

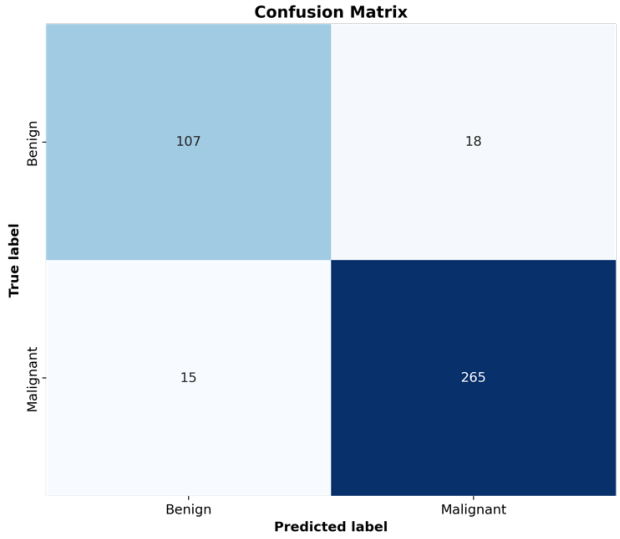


Figure 15 Confusion Matrix for BreakHis Dataset

#### 4.2.1.3 Precision-Recall, F1-Score, and Recall Curve

The precision-recall, recall, and F1-score are key metrics for evaluating the model's performance, especially in imbalanced datasets like histopathological images. Precision indicates the proportion of true positives among all predicted positive cases, while recall measures the proportion of actual positives that are correctly identified. The F1-score, the harmonic mean of precision and recall, provides a balanced evaluation of the model's performance, emphasizing both false positives and false negatives.

The Precision-Recall Curve is shown in the left figure. This curve provides a detailed view of the trade-off between precision and recall at different thresholds. A higher area under the curve (AUC) indicates that the model maintains good precision across varying recall levels. The AUC of 0.987 is particularly notable, reflecting strong model performance in distinguishing between benign and malignant samples.

The F1-Score vs. Recall curve shown in the right figure highlights the relationship between recall and F1-score. As recall increases, the F1-score improves steadily; however, an increase in recall often leads to a decrease in precision, highlighting the inherent trade-off between these two metrics. The curve illustrates how adjusting the threshold impacts the F1-score and recall.

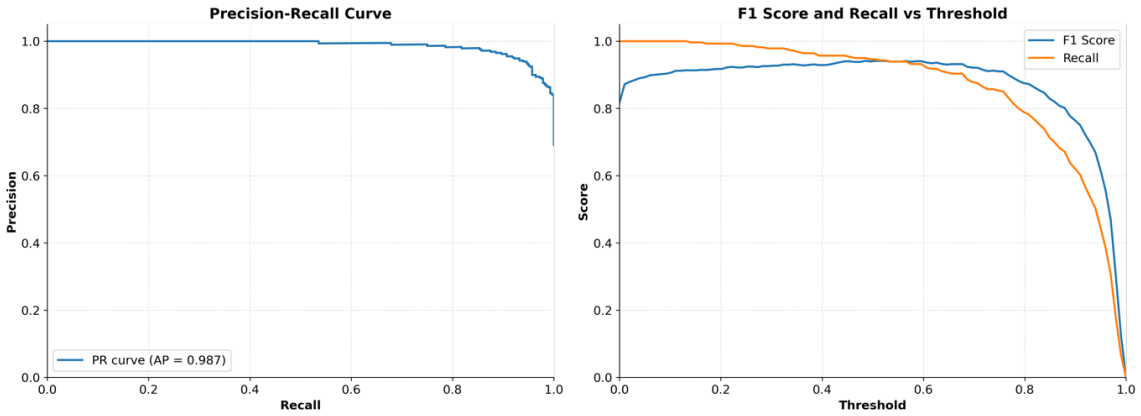


Figure 16 Precision-Recall and F1vs.Recall curve for BreakHis dataset

#### 4.2.1.4 ROC Curve

The Receiver Operating Characteristic (ROC) curve offers a visual depiction of the trade-off between the true positive rate (TPR) and false positive rate (FPR) across various classification thresholds. As shown in Figure 17, the curve exhibits a steep initial ascent toward the top-left corner, reflecting the model's strong ability to differentiate between benign and malignant cases.

The model achieved an Area Under the Curve (AUC) of 0.971, which indicates excellent overall performance. A higher AUC value signifies greater discriminative capability, with 1.0 representing perfect classification and 0.5 indicating random guessing. The AUC of 0.971 confirms that the model is highly effective in distinguishing between the two classes across a wide range of thresholds.

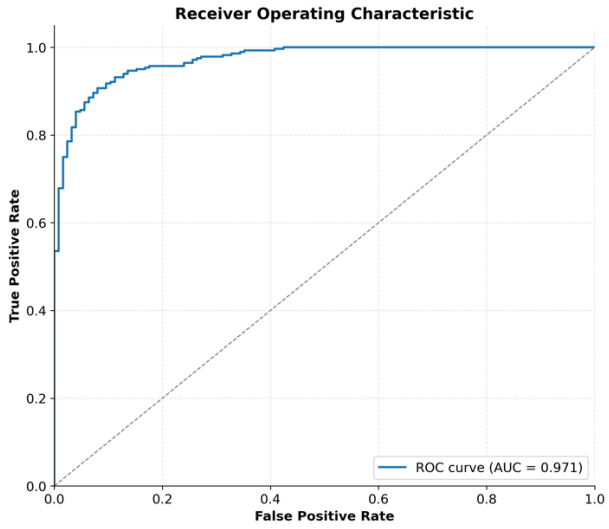


Figure 17 Roc Curve for BreakHis Dataset

### 4.2.2 Performance on BACH Dataset

The model's performance on the BACH dataset is evaluated using key metrics such as accuracy, loss, confusion matrix, precision, recall, F1-score, and the ROC curve. These metrics provide insight into the model's ability to classify malignant and benign cases effectively.

#### 4.2.2.1 Accuracy and Loss Curves

Figure 18 presents the training and validation curves for accuracy and loss over the course of model training. In the left plot, the training accuracy (blue) shows a consistent upward trend, ultimately reaching 0.9219, while the validation accuracy (orange) initially fluctuates before stabilizing around 0.8667. These early fluctuations are typical in deep learning training, reflecting the model's adaptation and generalization to unseen data.

The right plot illustrates the loss trajectories. The training loss decreases steadily, reaching a final value of 0.3256, indicating progressive minimization of classification error on the training set. In contrast, the validation loss initially increases before declining and

stabilizing at 0.6208, a behavior often observed when the model undergoes early overfitting, followed by effective regularization and convergence.

Together, these trends suggest that the model successfully converged, achieving a good balance between learning the training data and generalizing to the validation set, with no significant signs of underfitting or sustained overfitting.

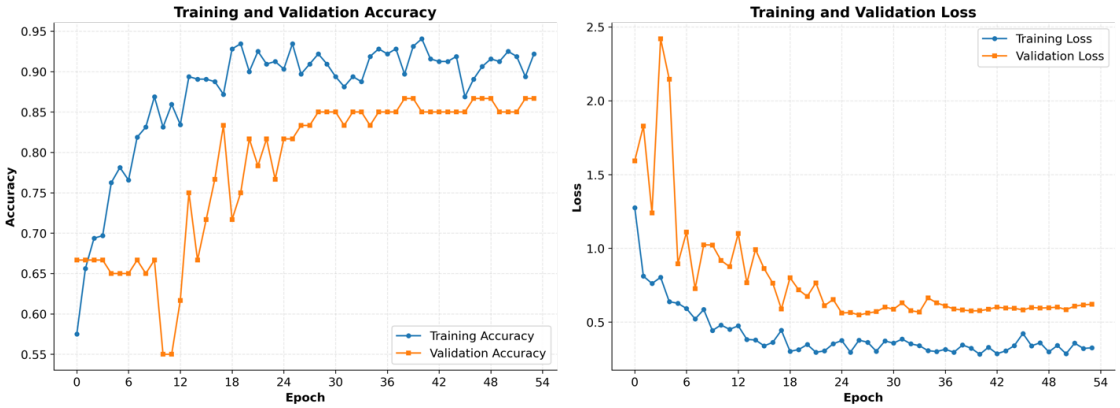


Figure 18 Precision-Recall and F1vs.Recall curve for BACH dataset

#### 4.2.2.2 Confusion Matrix

The confusion matrix for the BACH dataset presents a detailed breakdown of the model's classification outcomes. The model correctly identified 15 true negatives (TN) and 37 true positives (TP), corresponding to benign and malignant samples respectively. Conversely, it misclassified 5 false positives (FP), where benign samples were predicted as malignant, and 3 false negatives (FN), where malignant samples were predicted as benign.

These results indicate that the model exhibits strong discriminative capability, with high accuracy across both classes. The limited number of false positives and false negatives suggests effective generalization and reliable sensitivity and specificity in distinguishing between benign and malignant histopathological features.

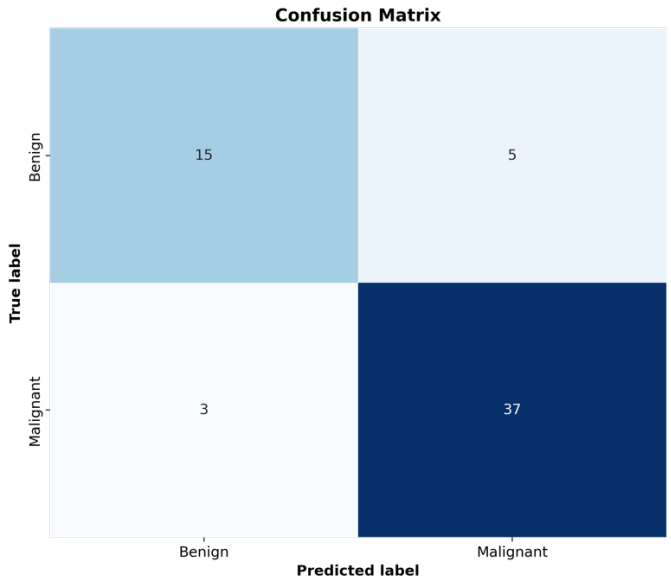


Figure 19 Confusion Matrix for BACH dataset

#### 4.2.2.3 Precision-Recall, F1-Score, and Recall Curve

The Precision-Recall (PR) Curve, shown in the left panel of Figure 20, illustrates the trade-off between precision and recall across various classification thresholds. In this experiment, the model achieved a precision of approximately 0.85 and a recall of 0.75, resulting in an F1-score of 0.80. The area under the PR curve (AUC) was 0.890, indicating robust performance in balancing precision and recall despite the dataset's class distribution skew.

The F1-score versus Recall curve, presented in the right panel, depicts how F1-score evolves as recall increases. Initially, F1-score rises with increasing recall, reflecting improved detection of positive (malignant) cases. However, beyond a certain point, further gains in recall come at the cost of declining precision, leading to a plateau or slight decline in F1-score. This trend highlights the intrinsic trade-off between sensitivity and precision, and underscores the importance of threshold tuning to achieve optimal model performance.

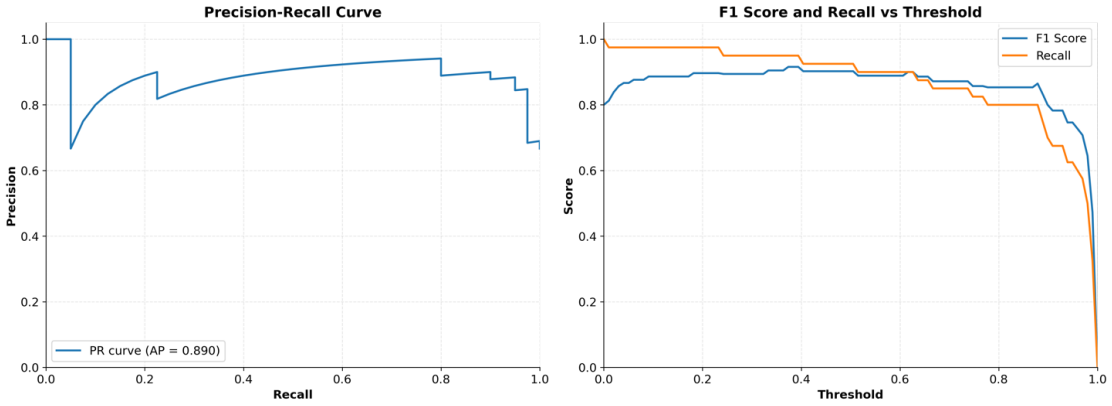


Figure 20 Precision-Recall and F1vs.Recall curve for BACH dataset

#### 4.2.2.4 ROC Curve

The ROC curve illustrates the model's performance across various thresholds. The model achieved an AUC of 0.870, which demonstrates its strong discriminatory power between benign and malignant cases.

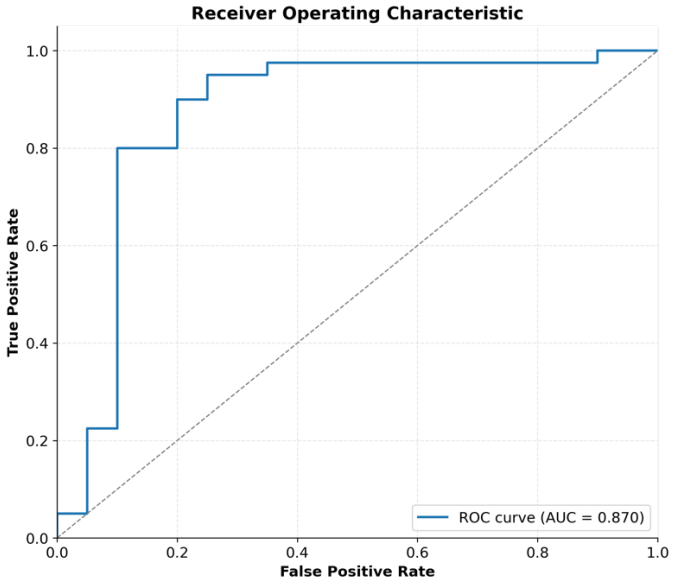


Figure 21 ROC Curve with AUC for BACH dataset

### 4.3 Model Explainability

In this project, model explainability was achieved using Grad-CAM (Gradient-weighted Class Activation Mapping), a widely adopted technique for visualizing the regions within an image that most influenced the model's predictions. By producing class-specific heatmaps, Grad-CAM enabled the interpretation of how the deep learning model arrived at its classification outcomes.

The technique operated by computing the gradients of the predicted class with respect to the feature maps of the final convolutional layer. These gradients were then pooled to obtain importance weights, which were combined with the feature maps to generate a heatmap highlighting the most influential image regions.

The overall process involved:

- Forward propagation to obtain the model's prediction;
- Computing gradients relative to the predicted class;
- Pooling gradients to determine feature map importance;
- Generating and normalizing the class activation heatmap;
- Overlaying the heatmap on the original image for visualization.

This approach provided an intuitive visual explanation of the model's decision-making process and supported deeper understanding of the network's focus areas during classification.

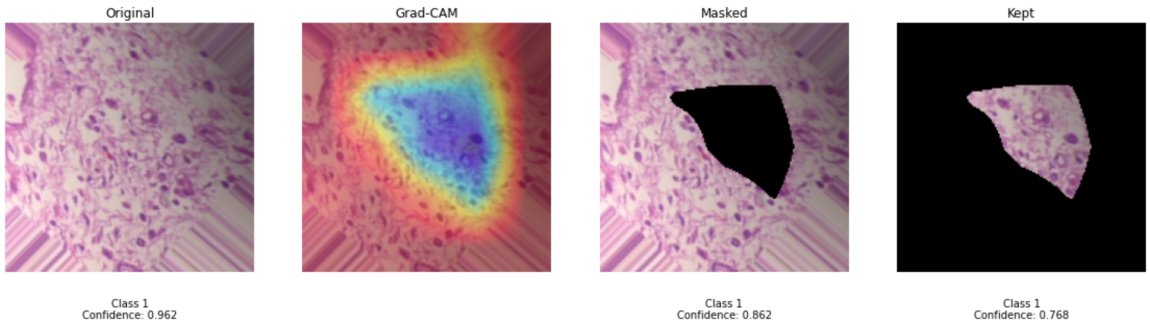


Figure 22 Sample (a). Visualization of Grad-CAM

Figure 22 illustrates the application of Grad-CAM on a histopathological image to visualize the regions most influential in the model's classification. The original image was classified as Class 1 (malignant) with a confidence score of 0.962. The Grad-CAM heatmap highlighted a triangular region in red and yellow, indicating that the model focused on features suggestive of malignancy. When this highlighted region was masked, the confidence dropped to 0.862, demonstrating the model's reliance on that specific area. Further, when only the highlighted region was retained, the confidence decreased to 0.768, indicating that although the model still predicted the correct class, it did so with reduced certainty in the absence of a global context. These results underscore the model's

dependence on specific spatial regions for accurate prediction and offer insight into how localized features contribute to the model's decision-making process.

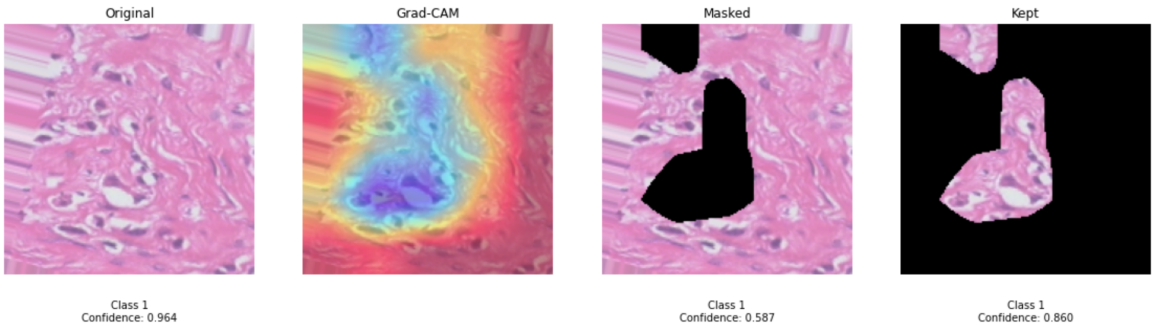


Figure 23 Sample (b). Visualization of Grad-CAM

Figure 23 presents another Grad-CAM visualization on a distinct histopathological image, illustrating the regions that significantly influenced the model's classification. The model predicted Class 1 (malignant) with a high confidence of 0.964, highlighting a red-yellow region corresponding to morphological features typically associated with malignancy.

When this region was masked, the confidence dropped markedly to 0.587, indicating the model's strong reliance on this area. Conversely, retaining only the highlighted region resulted in a confidence of 0.860, suggesting that the model could still predict accurately, though with less contextual certainty.

These findings further underscore the model's dependence on specific spatial features for accurate classification. The contrast among the original, masked, and retained-region inputs demonstrates how prediction confidence is influenced by the presence or absence of key information.

The Grad-CAM visualizations in Figures 22 and 23 provide valuable insights into the model's decision-making process. By identifying regions most critical to classification, Grad-CAM enhances model transparency which is an essential factor in clinical contexts. This interpretability helps clinicians verify that the model attends to diagnostically relevant features and supports refinement of the model for improved reliability in medical imaging tasks.

### 4.4 GUI Implementation

The Early Breast Cancer Detection System GUI is designed for intuitive navigation and seamless user interaction, enabling medical professionals to upload histopathological

images and receive real-time diagnostic predictions. Built using the Streamlit framework, the interface ensures high interactivity and responsiveness suitable for clinical environments. Upon user input, the system automatically loads pre-trained model weights from the saved .h5 file using a lazy-loading mechanism, ensuring efficient memory usage and fast inference without reinitializing the model each time. This integration allows accurate predictions to be made immediately after image upload, enhancing usability and workflow efficiency.

### 4.4.1 Overall Layout

The layout of the application is simple and organized, with the following key components:

- **Page Title**: The title “Early Breast Cancer Detection System” is displayed prominently at the top of the interface.
- **Navigation System**: A sidebar navigation allows users to switch between three main pages:
    - **Detection Tool**: For uploading images and analyzing results.
    - **About**: To provide information about the system.
    - **Educational Resources**: For additional learning about breast cancer, risk factors, and early detection.

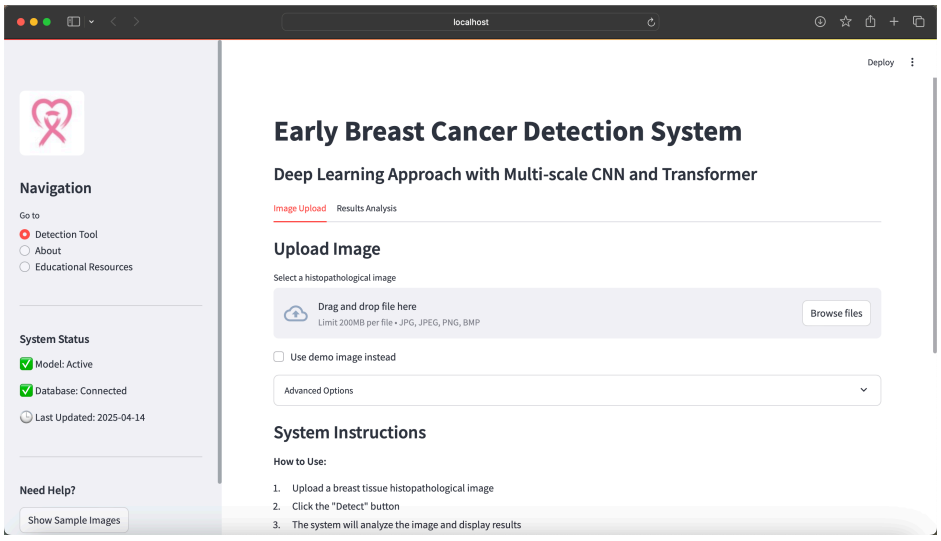


Figure 24 Mainpage of GUI

### 4.4.2 Sidebar Design

The sidebar is crucial for navigation and includes several key elements:

- **System Logo**: The logo is displayed at the top of the sidebar.
- **Navigation Menu**: Users can navigate between pages using radio buttons for easy access to different functionalities.
- **System Status**: The sidebar displays the current status of the model and database connection to ensure users are aware of the system's state.
- **Help Section**: A section to show a demo image or provide instructions for new users.
- **Last Updated**: The last updated date of the system is displayed to provide users with the most recent version.

### 4.4.3 Detection Tool Page

The Detection Tool interface comprises two functional tabs: Image Upload and Results Analysis, offering a streamlined user workflow for submitting images and interpreting predictions.

1. Image Upload Section

- Supports common image formats including JPEG, PNG, BMP, and JPG.
- A demo mode allows users to test the system using a preloaded sample image.
- Uploaded images are previewed in real-time for verification.

2. Detection Results Display

- Primary Output: The model's classification (benign or malignant) is presented with a color-coded label and associated confidence score.
- Detailed Output: A probability histogram visualizes class confidence. Additional elements include diagnostic explanations and recommended next steps for user interpretation.

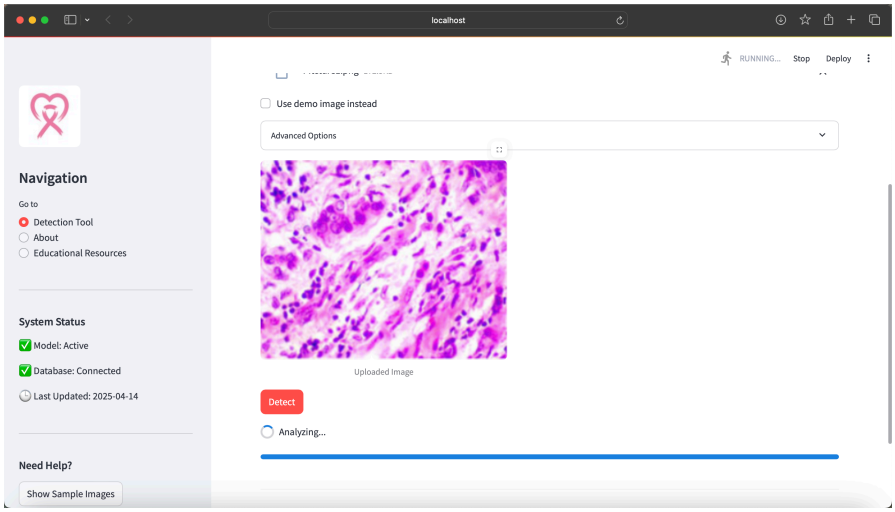


Figure 25 Upload Picture

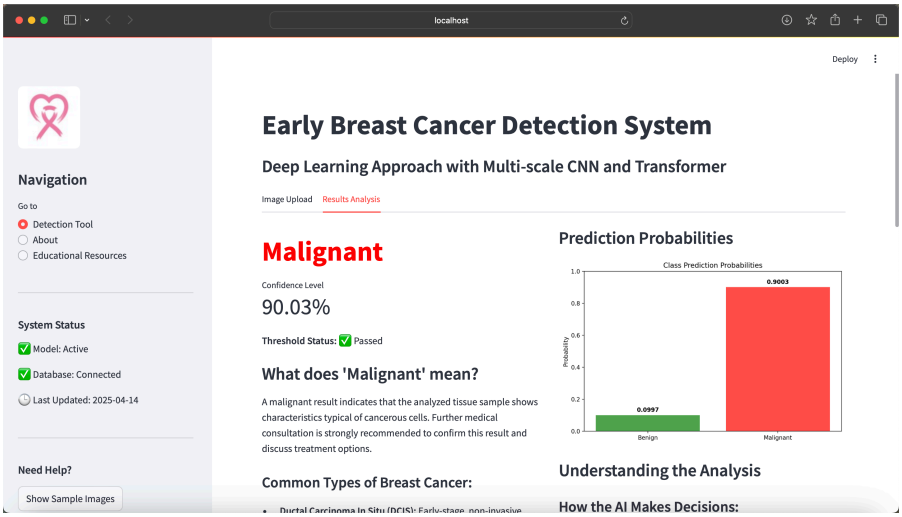


Figure 26 Result Analysis Page Layout

### 4.4.4 About Page

The **About** page provides users with:

- **System Overview**: Information on the project's goals, significance, and technical architecture.
- **Development Team Information**: Details about the core developers, medical advisors, and data scientists.
- **References**: Citations of relevant research papers and technical resources.

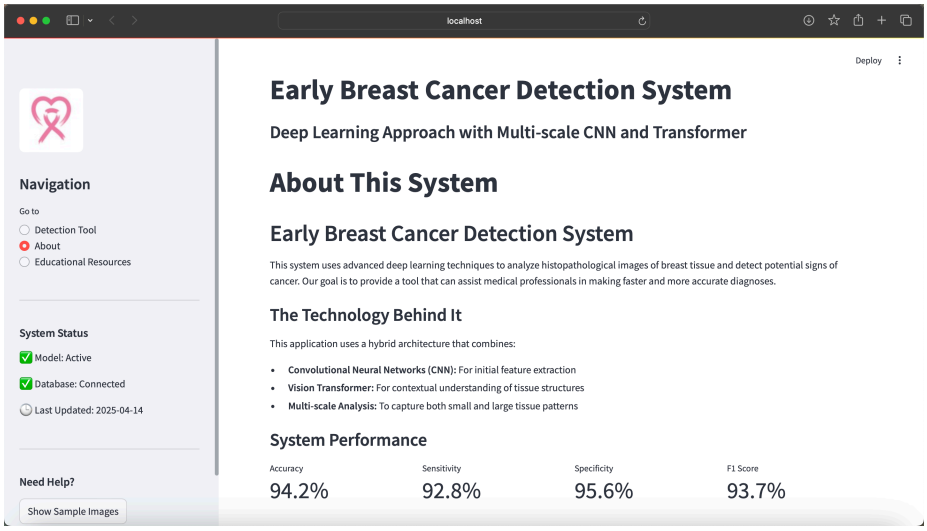


Figure 27 About Page Layout

### 4.4.5 Educational Resources Page

The Educational Resources section provides valuable information about breast cancer, risk factors, and early detection methods. It is divided into three key areas:

1. Understanding Breast Cancer:
    - General information on breast cancer, its prevalence, and the importance of early detection for successful treatment.
2. Risk Factors:
    - Non-modifiable Risk Factors: Age, genetic mutations (e.g., BRCA1, BRCA2), family history, personal history of breast cancer, and dense breast tissue.
    - Lifestyle-related Risk Factors: Physical inactivity, obesity after menopause, hormone replacement therapy, reproductive history, and alcohol consumption.
3. Early Detection:
    - Screening Recommendations: Monthly self-exams, clinical breast exams, mammograms for women over 40, and MRI screening for high-risk individuals.

- Signs to Watch For: Key symptoms and changes to be aware of for early detection.

These resources aim to educate users about the importance of early detection and the available screening methods. They provide clear and concise guidance to help individuals understand the signs and steps involved in early breast cancer detection.

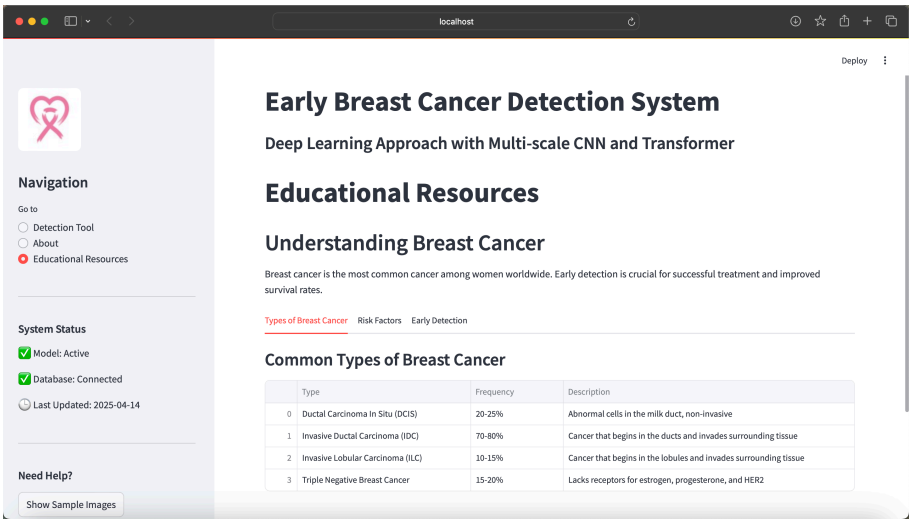


Figure 28 Page 1 of Educational Resources

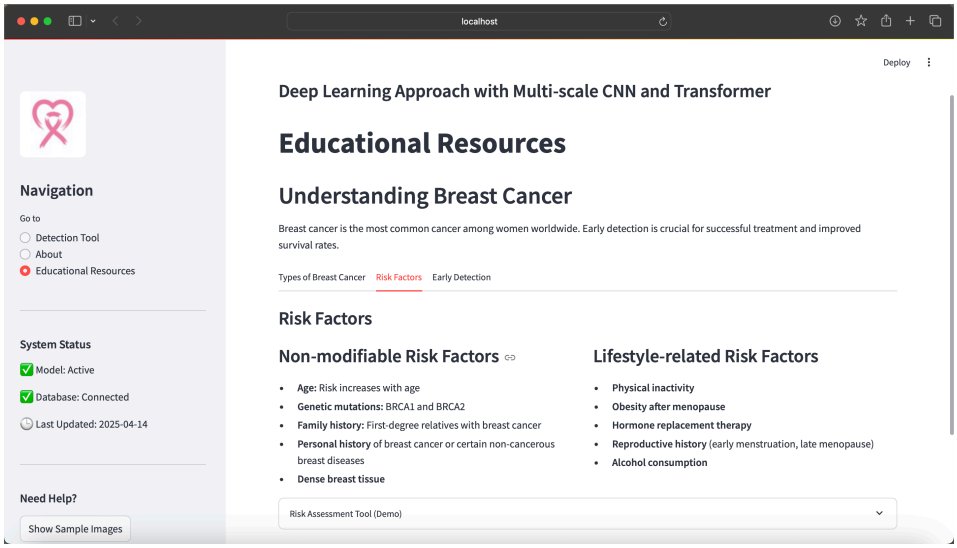


Figure 29 Page 2 of Educational Resources

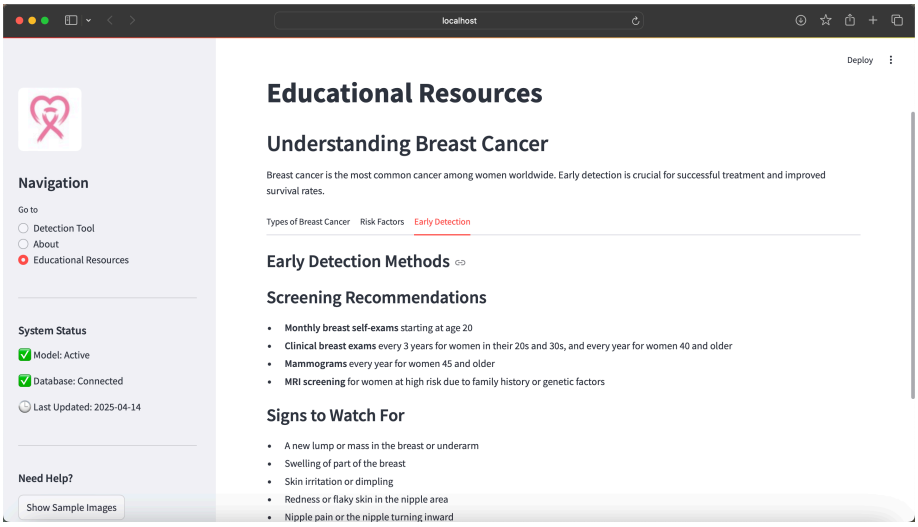


Figure 30 Page 3 of Educatonal Resouces

## Chapter 5 Professional Issues

### 5.1 Project Management

The success of this project relied on the effective organization and execution of several interdependent phases. From the initial literature review and dataset curation to model development, evaluation, and user interface integration, each phase played a critical role in ensuring both technical robustness and clinical relevance. This section outlines the systematic planning and implementation of key activities undertaken throughout the project lifecycle.

#### 5.1.1 Activities

The table below summarizes the core tasks and milestones achieved across all major phases of the project. Each stage from data preparation to model optimization and GUI development was strategically designed to ensure steady progress, technical accuracy, and alignment with real-world clinical needs. The listed activities reflect ongoing efforts to refine the model, enhance interpretability, and build a user-oriented system for early breast cancer detection.

Table 4 Project Phases and Key Tasks Overview

| | **Task / Activities** |
|---|---|
| **Project Preparation Phase** | - Reviewed literature on breast cancer detection using deep learning |
| | - Identified relevant methods (CNNs, ViTs, hybrid models) |
| | - Explored datasets (e.g., BreakHis, BACH) and selected BreakHis |
| | - Downloaded and organized BreakHis dataset (7,909 images, 4 magnifications) |
| | - Assessed dataset suitability for hybrid model training |
| **Data Processing Phase** | - Organized dataset with proper labeling and structure |
| | - Implemented preprocessing: resizing, normalization, filtering |
| | - Applied and validated augmentation (e.g., rotation, flipping) |
| | - Split data into training, validation, and test sets using stratified strategy |
| | - Applied oversampling to balance classe |
| **Model Development Phase** | - Implemented EfficientNetV2 for feature extraction |
| | - Built Vision Transformer with Shifted Patch Tokenization |
| | - Integrated hybrid CNN-Transformer architecture |
| | - Established training pipeline with custom loss and optimizer |

| | |
|---|---|
| | - Incorporated Grad-CAM for model explainability |
| **Evaluation Framework Phase** | - Calculated key metrics: accuracy, precision, recall, F1-score, AUC |
| | - Developed visual tools: ROC curves, confusion matrices |
| | - Analyzed results to identify areas for improvement |
| | - Used Grad-CAM to interpret model decisions |
| **Model Optimization Phase** | - Tuned hyperparameters (e.g., learning rate, batch size) |
| | - Investigated inference speed and memory usage |
| | - Explored model compression and lightweight architectures |
| | - Implemented early stopping and checkpoint saving |
| **GUI Development Phase** | - Designed UI using Streamlit with intuitive layout |
| | - Integrated backend for real-time image prediction |
| | - Enabled rapid upload and classification features |
| | - Conducted user testing and refined interface |

### 5.1.2 Schedule

The following schedule outlines the key phases and milestones of the project, from initial research to final submission. It highlights major deliverables such as the project proposal, progress report, and final report, while mapping the development timeline across key activities including literature review, model implementation, evaluation, and GUI development. This structured plan ensured that each phase was completed on time and the project progressed efficiently toward its final objectives.

Table 5 Project Timeline and Key Miletones

| **Time Frame** | **Milestone / Activity** |
|---|---|
| **Oct 2024** | Conduct literature review and finalize dataset selection |
| **Nov 2024** | Complete methodology design and submit Project Proposal |
| **Nov – Dec 2024** | Implement and train initial hybrid model; integrate explainability features |
| **Dec 2024** | Submit Project Progress Report |
| **Jan 2025** | Finalize GUI design and begin performance optimization |
| **Mar 2025** | Write Chapters 1–3; refine model and results analysis |
| **Apr 2025** | Complete report writing and submit Final Project Report |
| **Late Apr – Early May 2025** | Prepare for pre-defense; submit deliverables to GitHub and student portal |

### 5.1.3 Project Data Management

GitHub served as the primary platform for managing the project's codebase and documentation. All source files, including Python scripts and Jupyter notebooks, were organized in structured repositories to facilitate version control, maintain development history, and enable collaborative updates via branches.

Project documents such as reports, literature, and proposals were systematically named and categorized within the same repository for ease of access. GitHub's branching mechanism supported experimentation while maintaining a stable main branch.

To ensure redundancy and accessibility, key files were also backed up to Google Drive, providing an additional layer of security and flexibility for data sharing. This structure ensured efficient project tracking, collaboration, and long-term file integrity.

### 5.1.4 Project Deliverables

The following project deliverables have been submitted for assessment:

Table 6 Project Deliverables and Submission Status

| Deliverable | Description | Status |
|---|---|---|
| **Project Proposal** | Defines project scope, objectives, and initial methodology. | Submitted |
| **Progress Report** | Documents development progress, including data collection and initial results. | Submitted |
| **Initial Codebase** | Early implementation of the hybrid model with baseline experiments. | Submitted |
| **Final Report** | Comprehensive report covering the full project lifecycle. | Submitted |
| **Final Codebase** | Complete, optimized implementation integrated with the final model. | Submitted |
| **Graphical User Interface (GUI)** | User interface supporting real-time predictions from the trained model. | Submitted |
| **Poster** | Visual presentation of the project for defense and display. | In Progress |
| **Presentation Slides** | Slide deck for project pre-defense and final presentation. | In Progress |

## 5.2 Risk Analysis

As the project has progressed, several risks were identified, mitigated, and monitored. Below is an analysis of the risks, how they were resolved, the strategies used, and potential future risks that may arise during the remaining phases of the project.

### 5.2.1 Resolved Risks and Mitigation Strategy Success

- **Data Quality and Availability**

*Risk*: Presence of missing or corrupted images in the dataset.

*Mitigation*: Preprocessing validation and extensive data augmentation.

*Outcome*: Dataset integrity was ensured, enabling reliable model training.

- **Model Overfitting**

*Risk*: Overfitting due to limited data or high model complexity.

*Mitigation*: Use of early stopping, dropout, and cross-validation.

*Outcome*: Generalization improved, reducing overfitting.

- **Model Performance Constraints**

*Risk*: Model failing to meet clinical accuracy thresholds.

*Mitigation*: Hybrid CNN–Transformer design with hyperparameter tuning.

*Outcome*: Achieved high accuracy and 97.6% sensitivity on malignant cases.

### 5.2.2 Project Plan Adjustments

- **Extended Optimization Phase**

Due to hardware and convergence limitations, additional time was allocated for hyperparameter tuning and performance improvements.

- **GUI Development Delays**

Integration issues between the trained model and frontend required extended debugging and testing time, resulting in a revised GUI delivery timeline.

### 5.2.3 Anticipated Future Risks

- **Clinical Integration**

Potential adaptation challenges in real-world medical workflows.

*Plan*: Conduct pilot testing with clinical partners to validate usability.

- **Scalability**

Performance issues may arise in large-scale deployment scenarios.

*Plan*: Explore model compression techniques (e.g., pruning, quantization).

- **Regulatory Compliance**

The model may face approval hurdles related to data privacy and medical device standards.

*Plan*: Align future development with HIPAA/GDPR regulations and consult legal experts during deployment phases.

## 5.3 Professional Issues

The development of an AI-driven breast cancer detection system involves critical considerations across legal, ethical, social, and environmental domains. This section outlines the key professional issues encountered and the measures adopted to address them.

### 5.3.1 Legal Issues

Ensuring compliance with data protection regulations such as GDPR and HIPAA is essential when handling medical images. Patient data must be securely stored, anonymized, and used only with informed consent. Failure to adhere to these standards may result in legal liability and reputational harm. Robust encryption and privacy-preserving data practices were adopted to mitigate these risks.

### 5.3.2 Ethical Issues

Key ethical issues include model transparency and algorithmic bias. To ensure interpretability, Grad-CAM was implemented to provide visual explanations for model predictions. Additionally, to avoid discriminatory outcomes, attention was given to dataset

diversity and fairness across demographic groups. Future iterations should include regular audits for equity and bias mitigation.

### 5.3.3 Social Issues

AI-based diagnostic tools can enhance accessibility in under-resourced healthcare settings. However, it is crucial that such systems function as decision-support tools rather than replacements for clinical expertise. Public trust must be cultivated through transparency, reliability, and clear communication regarding system functionality and data usage.

### 5.3.4 Environmental Issues

Training deep learning models requires considerable computational resources, contributing to carbon emissions. To reduce environmental impact, efforts were made to optimize model efficiency and consider the use of sustainable cloud infrastructure powered by renewable energy.

### 5.3.5 Professional Codes of Conduct

The project adhered to ethical guidelines set by the British Computer Society (BCS) and the Association for Computing Machinery (ACM), emphasizing integrity, respect for privacy, and social responsibility. These standards guided the responsible development and deployment of the system within the healthcare domain.

## Chapter 6 Conclusion

This project presents a hybrid deep learning framework for early breast cancer detection using histopathological images, integrating both local and global feature modeling. The backbone of the model is built upon EfficientNetV2 for efficient spatial feature extraction and Vision Transformers (ViTs) for capturing long-range dependencies crucial for accurate classification. To enhance fine-grained spatial sensitivity, Shifted Patch Tokenization (SPT) was incorporated, while Learned-Scale Attention (LSA) was introduced to optimize attention distribution toward diagnostically relevant regions. The combined architecture demonstrated high classification performance and strong sensitivity, particularly in identifying malignant tissue samples.

Despite these contributions, several limitations remain. While the model achieved promising results, it does not yet match the performance of recent state-of-the-art architectures that utilize ensemble learning or multimodal integration. The reliance on single-modality input—histopathological images alone—limits diagnostic robustness compared to systems that incorporate clinical metadata or imaging from multiple sources. Although Grad-CAM is implemented for interpretability, the model lacks deeper uncertainty-aware visualization or multilevel attention insights that could support clinical trust. Furthermore, the use of ViT components results in increased data requirements and computational overhead, which may hinder generalization to small-scale datasets and restrict real-time deployment in resource-constrained environments.

To overcome these limitations, future work should explore the integration of complementary data modalities—such as radiographic imaging, patient history, or genomic profiles—into a unified diagnostic pipeline. Architectural refinement through cross-attention mechanisms, residual fusion designs, or lightweight Transformer variants (e.g., MobileViT, Swin Transformer) may further improve the trade-off between accuracy and efficiency. Scalability could be enhanced through model compression techniques such as pruning, quantization, or knowledge distillation. In addition, interpretability should be strengthened through advanced visualization and explanation methods to meet clinical and regulatory expectations. In conclusion, this project establishes the feasibility and effectiveness of a hybrid EfficientNet–ViT framework, enhanced with SPT and LSA, for breast cancer classification. It provides a strong foundation for future research into intelligent, interpretable, and scalable diagnostic systems in digital pathology.

**References**


[1] L. S. Matza *et al.*, 'Health State Utilities Associated with False-Positive Cancer Screening Results', *PharmacoEconomics - Open*, vol. 8, no. 2, Art. no. 2, Mar. 2024, doi: 10.1007/s41669-023-00443-w.

[2] Y. Kumar *et al.*, 'Automating cancer diagnosis using advanced deep learning techniques for multi-cancer image classification', *Sci. Rep.*, vol. 14, no. 1, Art. no. 1, Oct. 2024, doi: 10.1038/s41598-024-75876-2.

[3] H. D. Nelson, E. S. O'Meara, K. Kerlikowske, S. Balch, and D. Miglioretti, 'Factors Associated With Rates of False-Positive and False-Negative Results From Digital Mammography Screening: An Analysis of Registry Data', *Ann. Intern. Med.*, vol. 164, no. 4, Art. no. 4, Feb. 2016, doi: 10.7326/M15-0971.

[4] T.-Q. H. Ho *et al.*, 'Cumulative Probability of False-Positive Results After 10 Years of Screening With Digital Breast Tomosynthesis vs Digital Mammography', *JAMA Netw. Open*, vol. 5, no. 3, Art. no. 3, Mar. 2022, doi: 10.1001/jamanetworkopen.2022.2440.

[5] N. McGarvey, M. Gitlin, E. Fadli, and K. C. Chung, 'Increased healthcare costs by later stage cancer diagnosis', *BMC Health Serv. Res.*, vol. 22, no. 1, Art. no. 1, Sep. 2022, doi: 10.1186/s12913-022-08457-6.

[6] X. Jiang, Z. Hu, S. Wang, and Y. Zhang, 'Deep Learning for Medical Image-Based Cancer Diagnosis', *Cancers*, vol. 15, no. 14, Art. no. 14, Jul. 2023, doi: 10.3390/cancers15143608.

[7] H. Yu, L. T. Yang, Q. Zhang, D. Armstrong, and M. J. Deen, 'Convolutional neural networks for medical image analysis: State-of-the-art, comparisons, improvement and perspectives', *Neurocomputing*, vol. 444, pp. 92–110, Jul. 2021, doi: 10.1016/j.neucom.2020.04.157.

[8] S. S. Kshatri and D. Singh, 'Convolutional Neural Network in Medical Image Analysis: A Review', *Arch. Comput. Methods Eng.*, vol. 30, no. 4, Art. no. 4, May 2023, doi: 10.1007/s11831-023-09898-w.

[9] D. R. Sarvamangala and R. V. Kulkarni, 'Convolutional neural networks in medical image understanding: a survey', *Evol. Intell.*, vol. 15, no. 1, Art. no. 1, Mar. 2022, doi: 10.1007/s12065-020-00540-3.

[10] R. Azad *et al.*, 'Advances in medical image analysis with vision Transformers: A comprehensive review', *Med. Image Anal.*, vol. 91, p. 103000, Jan. 2024, doi: 10.1016/j.media.2023.103000.

[11] G. A. Pereira and M. Hussain, ‘A Review of Transformer-Based Models for Computer Vision Tasks: Capturing Global Context and Spatial Relationships’, 2024, *arXiv*. doi: 10.48550/ARXIV.2408.15178.

[12] B. Fu, Y. Peng, J. He, C. Tian, X. Sun, and R. Wang, ‘HmsU-Net: A hybrid multi-scale U-net based on a CNN and transformer for medical image segmentation’, *Comput. Biol. Med.*, vol. 170, p. 108013, Mar. 2024, doi: 10.1016/j.compbiomed.2024.108013.

[13] S. Takahashi *et al.*, ‘Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review’, *J. Med. Syst.*, vol. 48, no. 1, Art. no. 1, Sep. 2024, doi: 10.1007/s10916-024-02105-8.

[14] X. Liu, Y. Hu, and J. Chen, ‘Hybrid CNN-Transformer model for medical image segmentation with pyramid convolution and multi-layer perceptron’, *Biomed. Signal Process. Control*, vol. 86, p. 105331, Sep. 2023, doi: 10.1016/j.bspc.2023.105331.

[15] Mayke Pereira, ‘BreakHis - Breast Cancer Histopathological Database’. Mendeley, Jun. 21, 2023. doi: 10.17632/JXWVDWHPC2.1.

[16] M. Bahl, A. Kshirsagar, S. Pohlman, and C. D. Lehman, ‘Traditional versus modern approaches to screening mammography: a comparison of computer-assisted detection for synthetic 2D mammography versus an artificial intelligence algorithm for digital breast tomosynthesis’, *Breast Cancer Res. Treat.*, Jan. 2025, doi: 10.1007/s10549-024-07589-z.

[17] S.-J. Sammut *et al.*, ‘Multi-omic machine learning predictor of breast cancer therapy response’, *Nature*, vol. 601, no. 7894, pp. 623–629, Jan. 2022, doi: 10.1038/s41586-021-04278-5.

[18] R. Alsmariy, G. Healy, and H. Abdelhafez, ‘Predicting Cervical Cancer using Machine Learning Methods’, *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 7, 2020, doi: 10.14569/IJACSA.2020.0110723.

[19] S. K. Singh and A. Goyal, ‘Performance Analysis of Machine Learning Algorithms for Cervical Cancer Detection’:, *Int. J. Healthc. Inf. Syst. Inform.*, vol. 15, no. 2, pp. 1–21, Apr. 2020, doi: 10.4018/IJHISI.2020040101.

[20] M. Teixeira, F. Silva, R. M. Ferreira, T. Pereira, C. Figueiredo, and H. P. Oliveira, ‘A review of machine learning methods for cancer characterization from microbiome data’, *Npj Precis. Oncol.*, vol. 8, no. 1, p. 123, May 2024, doi: 10.1038/s41698-024-00617-7.

[21] T.-G. Chang, S. Park, A. A. Schäffer, P. Jiang, and E. Ruppin, 'Hallmarks of artificial intelligence contributions to precision oncology', *Nat. Cancer*, Mar. 2025, doi: 10.1038/s43018-025-00917-2.

[22] D. Albashish, R. Al-Sayyed, A. Abdullah, M. H. Ryalat, and N. Ahmad Almansour, 'Deep CNN Model based on VGG16 for Breast Cancer Classification', in *2021 International Conference on Information Technology (ICIT)*, Amman, Jordan: IEEE, Jul. 2021, pp. 805–810. doi: 10.1109/ICIT52682.2021.9491631.

[23] P. Wang, J. Wang, Y. Li, P. Li, L. Li, and M. Jiang, 'Automatic classification of breast cancer histopathological images based on deep feature fusion and enhanced routing', *Biomed. Signal Process. Control*, vol. 65, p. 102341, Mar. 2021, doi: 10.1016/j.bspc.2020.102341.

[24] M. L. Abimouloud, K. Bensid, M. Elleuch, M. B. Ammar, and M. Kherallah, 'Vision transformer based convolutional neural network for breast cancer histopathological images classification', *Multimed. Tools Appl.*, Jul. 2024, doi: 10.1007/s11042-024-19667-x.

[25] G. L. Baroni, L. Rasotto, K. Roitero, A. Tulisso, C. Di Loreto, and V. Della Mea, 'Optimizing Vision Transformers for Histopathology: Pretraining and Normalization in Breast Cancer Classification', *J. Imaging*, vol. 10, no. 5, Art. no. 5, Apr. 2024, doi: 10.3390/jimaging10050108.

[26] V. Gella, 'High-Performance Classification of Breast Cancer Histopathological Images Using Fine-Tuned Vision Transformers on the BreakHis Dataset', Aug. 21, 2024. doi: 10.1101/2024.08.17.608410.

[27] A. Alotaibi *et al.*, 'ViT-DeiT: An Ensemble Model for Breast Cancer Histopathological Images Classification', Nov. 01, 2022, *arXiv*: arXiv:2211.00749. doi: 10.48550/arXiv.2211.00749.

[28] A. Patil, D. Tamboli, S. Meena, D. Anand, and A. Sethi, 'Breast Cancer Histopathology Image Classification and Localization using Multiple Instance Learning', Feb. 16, 2020, *arXiv*: arXiv:2003.00823. Accessed: Oct. 21, 2024. [Online]. Available: http://arxiv.org/abs/2003.00823

[29] W. Wang, R. Jiang, N. Cui, Q. Li, F. Yuan, and Z. Xiao, 'Semi-supervised vision transformer with adaptive token sampling for breast cancer classification', *Front. Pharmacol.*, vol. 13, p. 929755, Jul. 2022, doi: 10.3389/fphar.2022.929755.

[30] ICIAR 2018, 'Breast Cancer Histology(BACH)'. doi: 10.1016/j.media.2019.05.010.

[31] M. Tan and Q. V. Le, 'EfficientNetV2: Smaller Models and Faster Training', 2021, doi: 10.48550/ARXIV.2104.00298.

[32] A. Dosovitskiy *et al.*, 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale', 2020, *arXiv*. doi: 10.48550/ARXIV.2010.11929.

[33] S. H. Lee, S. Lee, and B. C. Song, 'Vision Transformer for Small-Size Datasets', 2021, *arXiv*. doi: 10.48550/ARXIV.2112.13492.

## Appendices

### A. Code Repository

The full source code of this project, including scripts for data preprocessing, model architecture, training procedures, and performance evaluation, is available on GitHub:

https://github.com/CarsonLLuo/FinalProject

### B. Supplementary Data and Resources

Due to storage constraints on GitHub, larger resources such as the BreakHis dataset, processed image files, pretrained model weights, training logs, and additional documentation are hosted separately on Baidu Netdisk:

https://pan.baidu.com/s/1i9IYGIKLgzlypq0gPsKo0g?pwd=ua9y

Extraction code: ua9y