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

OXFORD
BROOKES
UNIVERSITY

UNDERGRADUATE PROJECT PROGESS REPORT

Project Title:	**Study on Early Cancer Detection Using Multi-Scale CNN and Transformer-Based Deep Learning Approaches**
Surname:	**Luo**
First Name:	**Xinyu**
Student Number:	**202118020329**
Supervisor Name:	**Dr. Grace U. Nneji**
Module Code:	**CHC 6096**
Module Name:	**Project**
Date Submitted:	**December 27th , 2024**

Table of Contents

1 Introduction

1.1 Background

Early detection of cancer remains one of the most critical challenges in modern healthcare, with its importance underscored by 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]. While traditional detection methods such as mammography and biopsy have been instrumental in cancer diagnosis, they face limitations including invasiveness, high costs, and notable rates of false positives and negatives [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. These limitations not only increase healthcare costs but also create unnecessary patient anxiety and, in some cases, delay critical treatments that could potentially impact patient outcomes [5].

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]. 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 inherent limitations in capturing long-range dependencies and global contextual information due to their restricted receptive fields, which can lead to suboptimal performance when analyzing complex, high-resolution medical images[9].

The emergence of Vision Transformers (ViTs) has introduced new possibilities in medical image analysis. 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]. CNNs excel at capturing local features and processing high-resolution images, while Transformers are adept at modeling 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]. The work builds upon the proven effectiveness of CNNs in medical image analysis and the emerging potential of Vision Transformers in capturing global image context.

The availability of comprehensive datasets like BreakHis (Breast Cancer Histopathological Database) [15], which contains microscopic images at various magnification levels, provides an opportunity to develop and validate such hybrid approaches. The diversity in magnification levels (40X, 100X, 200X, and 400X) allows for thorough testing of models across different scales of tissue analysis, making it particularly suitable for developing robust cancer detection systems. This multi-scale nature of the dataset aligns well with the proposed hybrid architecture's ability to analyze images at different levels of detail.

This research aims to leverage these technological advances to develop a more accurate and reliable system for early cancer detection, potentially reducing the rate of false positives and negatives that currently challenge traditional diagnostic methods. By combining the strengths of both CNNs and Transformers, we seek 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 research is to develop and implement an innovative hybrid deep learning framework that combines enhanced Vision Transformers (ViT) with CNN architectures for accurate breast cancer detection in histopathological images. Based on our current progress, we have successfully developed several key components that contribute to this aim:

First, we have implemented a Shifted Patch Tokenization (SPT) mechanism to enhance the feature extraction capabilities of the standard ViT model, improving its ability to capture local spatial information. Second, we have developed a hybrid architecture that integrates EfficientNetV2's CNN capabilities with ViT's attention mechanisms, leveraging the complementary strengths of both approaches. Third, we have implemented an advanced attention mechanism through Learned-Scale Attention (LSA), which introduces learnable temperature parameters to optimize attention score distribution.

Through these implementations, we aim to achieve superior performance in breast cancer classification across multiple magnification levels (40x/100x/200x/400x) using the BreakHis dataset, while maintaining model interpretability and clinical relevance. Our comprehensive evaluation framework, incorporating multiple performance metrics and visualization tools, ensures rigorous validation of the model's effectiveness in real-world medical applications.

1.3 Objectives

1. Dataset Processing and Enhancement
   - Acquire and organize the BreakHis breast cancer histopathological dataset
   - Implement comprehensive data preprocessing including normalization and standardization
   - Develop and implement data augmentation strategies including rotation, scaling, translation, and flipping
   - Create efficient data pipelines for handling multi-scale images (40x/100x/200x/400x magnifications)
2. Model Architecture Development
   - Design and implement a baseline Vision Transformer (ViT) architecture
   - Develop an innovative Shifted Patch Tokenization (SPT) mechanism to enhance feature extraction
   - Create a hybrid architecture integrating EfficientNetV2 with ViT
   - Implement an advanced Learned-Scale Attention (LSA) mechanism for optimized attention distribution
3. Training and Optimization Strategy
   - Establish a robust training pipeline with comprehensive logging and monitoring
   - Implement learning rate scheduling and early stopping mechanisms
   - Develop model checkpointing and state management systems
   - Optimize hyperparameters through systematic experimentation
   - Design and implement efficient gradient computation and backpropagation strategies
4. Performance Evaluation Framework
   - Develop a comprehensive evaluation system including accuracy, precision, recall, and F1 scores
   - Create visualization tools for training progress and model performance analysis
   - Implement comparative analysis tools for different model architectures
   - Analyze model performance across various image magnification levels
5. System Integration and Deployment
   - Create a flexible configuration system for model parameters and training settings
   - Implement model serialization and loading functionalities
   - Develop an end-to-end inference pipeline for practical applications

- Establish a robust error handling and logging system

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 the BreakHis dataset, a comprehensive collection of breast cancer histopathological images captured at multiple magnification levels (40x, 100x, 200x, and 400x). This multi-scale approach allows for a thorough evaluation of tissue characteristics at different levels of detail. The project implements extensive data augmentation techniques and custom data splitting strategies, ensuring reliable model training and evaluation.

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 state-of-the-art 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.

The project represents a significant step forward in the application of deep learning to medical image analysis, combining theoretical innovation with practical clinical relevance. The research aims to contribute to the advancement of automated medical diagnostic tools while maintaining high standards of accuracy and reliability in cancer detection.

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 modeling 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.

2 Background Review

2.1 Traditional Method of Breast Cancer

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. Clinical breast exams, ultrasound, and biopsy were also employed, each with their own inherent challenges such as invasiveness and high costs, as discussed by McGarvey et al. [5]. These methods, while critical in their time, often led to unnecessary patient anxiety and potential delays in treatment due to their inaccuracies.

2.2 Machine Learning Method of Breast Cancer

The advent of machine learning in the field of breast cancer detection has significantly advanced the automation and precision of the diagnostic process. Early machine learning approaches, as mentioned by Albashish et al. [16], included the use of Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN) for classifying breast cancer from histopathological images. These methods, while foundational, relied on manual feature extraction, which was labor-intensive and often less precise compared to the capabilities of deep learning models that would emerge later.

Machine learning has since evolved to encompass a wider range of algorithms and techniques. Random forests, a popular ensemble method, and naive Bayes, a probabilistic classifier, have also been utilized in the classification of breast cancer. These algorithms, along with artificial neural networks (ANNs) and extreme learning machines (ELMs), have contributed to the development of more sophisticated models that can handle the complexity of medical imaging data.

2.3 Deep Learning Method of Breast Cancer

Deep learning has since become a pivotal tool in medical image analysis, particularly in the context of breast cancer detection.

2.3.1 CNN Models

Convolutional Neural Networks (CNNs) have emerged as a powerful tool for learning hierarchical features from medical images. Jiang et al. [6] demonstrated the remarkable success of CNNs in identifying patterns within histopathological images. CNNs, as reviewed by Kshatri and Singh [8], are adept at capturing local features and processing high-resolution images, which is crucial for detecting early-stage cancer. The ability of CNNs to learn from raw pixel data, as shown by Albashish et al. [16], has significantly improved the accuracy of cancer detection compared to traditional machine learning methods.

The success of CNNs in cancer detection can also be attributed to their adaptability to different types of medical imaging data, including MRI, CT scans, and histopathological images. This versatility has made CNNs a cornerstone in the development of automated cancer detection systems, which are essential for improving the efficiency and effectiveness of cancer screening and diagnosis.

2.3.2 Hybrid CNN mmodels

To overcome the limitations of standalone CNNs, particularly in capturing long-range dependencies and global contextual information, hybrid models have been developed. Wang et al. [17] proposed a hybrid CNN-Capsule Network (CapsNet) model that integrates both convolutional and capsule features to enhance classification performance. Abimouloud et al. [18] developed a hybrid Vision Transformer-CNN model that leverages the self-attention mechanism of ViTs and the feature extraction capabilities of CNNs. This approach, as demonstrated by Baroni et al. [19] and Gella [20], has achieved state-of-the-art performance on breast cancer histopathological images. The ensemble model of Vision Transformer (ViT) and Data-Efficient Image Transformer (DeiT) proposed by Alotaibi et al. [20] further showcases the power of ViTs in improving classification accuracy and reliability. Patil et al. [21] introduced an attention-based Multiple Instance Learning (MIL) approach, which not only improved classification but also provided better localization of malignant regions. Wang et al. [22] combined semi-supervised learning with ViTs, applying adaptive token sampling to significantly enhance breast cancer classification. These hybrid models, as evidenced by the works cited, have set new benchmarks in the field of breast cancer detection.

Table 1 Summary Table of Background Review

Author	Datasets	Methods & Models	Results
Dheeb Albashish et al.[16]	BreaKHis	VGG16 for feature extraction with classifiers (RBF-SVM, Poly-SVM, KNN, Logistic Regression, NN)	RBF-SVM achieved 96% accuracy for binary classification, and 89.83% accuracy for multiclass classification at 40x magnification.
Pin Wang et al.[17]	BreaKHis	FE-BkCapsNet: a dual-channel network combining CNN and CapsNet features with enhanced routing	Achieved classification accuracy of 92.71% at 40x, 94.52% at 100x, 94.03% at 200x, and 93.54% at 400x magnifications.
Mouhamed Laid Abimouloud et al.	BreaKHis	Hybrid models combining Vision Transformer (ViT), Compact Convolution Transformers (CCT), and Mobile Vision Transformers (MVIT)	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.[18]	BACH, BRACS, AIDPATH	Vision Transformer (ViT) pretrained on ImageNet with color normalization and data augmentation	Achieved 0.91 accuracy on BACH, 0.74 on BRACS, and 0.92 on AIDPATH dataset for tumor classification
Venkat Gella[20]	BreaKHis	Fine-tuned Vision Transformer (ViT) with Ranger optimizer	Achieved an accuracy of 99.99%, precision of 99.98%, and recall of 99.99% for binary classification
Amira Alotaibi et al.[23]	BreakHis	Ensemble model of Vision Transformer (ViT) and Data-Efficient Image Transformer (DeiT)	Achieved 98.17% accuracy, 98.18% precision, 98.08% recall, and a 98.12% F1 score for multi-class classification
Abhijeet Patil et al.[21]	BreaKHis, BACH	Attention-based Multiple Instance Learning (A-MIL)	Achieved classification accuracy of 86.56% at 200x magnification and effective localization of malignant regions
Wei Wang et al.[22]	BreaKHis, BUSI	Semi-supervised Vision Transformer (ViT) with Adaptive Token Sampling (ATS)	Achieved 98.12% accuracy, 98.17% precision, 98.65% recall, and 98.41% F1-score on BreaKHis

3 Technical Progress

3.1 Approach

3.1.1 BreakHis Dataset

The Breast Cancer Histopathological Database (BreakHis) [15] was employed in this study, comprising 7,909 microscopic images of breast tumor 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 tumors, providing a comprehensive foundation for binary classification of breast cancer histopathological images

For the dataset organization, let D_{total} represent the complete dataset in Equation 1:

$$\mathrm{D}_{\mathrm{total}} = \mathrm{D}_{\mathrm{train}} \cup \mathrm{D}_{\mathrm{val}} \cup \mathrm{D}_{\mathrm{test}} \qquad (1)$$

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

To ensure robust model evaluation, we implemented a custom data split strategy that differs from the original dataset organization. Rather than using the predefined train/test split, we 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. For each subset D_i, we maintain class balance through equation 2:

$$\frac{|D_i^{benign}|}{|D_i^{benign}| + |D_i^{malignant}|} \approx \frac{|D_{total}^{benign}|}{|D_{total}|} \qquad (2)$$

To address class imbalance issues in the training set, we applied random oversampling techniques. The oversampling process can be expressed as in equation 3:

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

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

3.1.2 Data Processing

The data preprocessing pipeline was carefully designed to optimize the dataset for deep learning model training. Each image underwent several crucial preprocessing steps. First, the images were resized to a uniform dimension of 160×160 pixels, chosen to balance computational efficiency with preservation of important histopathological features. For an input image I, the normalization process can be expressed as in equation 4:

$$I_{normalized} = \frac{I - min(I)}{max(I) - min(I)} \tag{4}$$

To enhance model robustness and prevent overfitting, we implemented comprehensive data augmentation techniques. The augmentation transformations can be represented as a composition of functions in equation 5:

$$I_{augmented} = T_n \circ T_{n-1} \circ \ldots \circ T_1(I) \tag{5}$$

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 image $I \in R^{160\times\,160\,\times\,3}$ properly normalized to $[0,1]^{160\times160\times3}$ and augmented when necessary, while maintaining the integrity of the evaluation process through consistent normalization across all sets.

3.1.3 EfficientNet-based Convolutional Neural Network

This research implements an EfficientNetV2-B0 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 in equation 6:

$$f(i,j) = \sum_m \sum_n I(i+m, j+n) \cdot K(m,n) \tag{6}$$

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):

$$MBConv(X) = PWC\left(DWC\big(PWC(X)\big)\right) \tag{7}$$

where PWC represents pointwise convolution and DWC represents depthwise convolution.

2. Fused-MBConv:

$$FusedMBConv(X) = PWC\big(Conv(X)\big) \tag{8}$$

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

$$F_l = H_l(F_{l-1}) \tag{9}$$

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:

$$X_{out} = X_{in} + \phi\left(BN\left(Conv\left(BN\left(Conv(X_{in})\right)\right)\right)\right) \quad (10)$$

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

$$\phi(x) = x \cdot sigmoid(x) \quad (11)$$

For optimization, the model employs the Adam optimizer with an inverse time decay learning rate schedule in equation 12:

$$\alpha_t = \alpha_0 \cdot \frac{1}{1 + \beta t} \quad (12)$$

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

The loss function utilizes sparse categorical cross-entropy in equation 13:

$$L = -\sum_{c=1}^{M} y_c \log(\widehat{y_c}) \quad (13)$$

where y_c represents the true label, $\widehat{y_c}$ 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 in equation 14, 15 and 16:

$$C_{out} = \alpha \cdot C_{in} \quad (14)$$

$$H_{out} = \frac{H_{in}}{\rho} \tag{15}$$

$$W_{out} = \frac{W_{in}}{\rho} \tag{16}$$

where α represents the channel multiplier and ρ 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.

3.1.4 Vision Transformer (ViT) with Shifted Patch Tokenization

The Vision Transformer (ViT) architecture 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 in equation 17:

$$P = \{p_1, p_2, \dots, p_N\} \text{ where } N = \frac{HW}{P^2} \tag{17}$$

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

This research enhances the standard ViT architecture with Shifted Patch Tokenization (SPT), which creates multiple views of the input image through spatial shifting in equation 18:

$$I_{shifted} = \{I, I_{left-up}, I_{left-down}, I_{right-up}, I_{right-down}\} \tag{18}$$

Each patch is then flattened and linearly projected to create token embeddings. Position embeddings are added to maintain spatial information in equation 19:

$$z_0 = \left[x_{class}; x_p^1 E; x_p^2 E; \dots; x_p^N E\right] + E_{pos} \tag{19}$$

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 in equation 20:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{20}$$

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:

$$MSA(X) = Concat(head_1, \dots, head_h)W^O \tag{21}$$

$$head_i = Attention\left(XW_i^Q, XW_i^K, XW_i^V\right) \tag{22}$$

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

$$x' = MSA\big(LN(x)\big) + x \tag{23}$$

$$x'' = FFN\big(LN(x')\big) + x' \tag{24}$$

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

$$FFN(x) = W_2\sigma(W_1 x + b_1) + b_2 \tag{25}$$

where σ is the GELU activation function.

For optimization, the architecture employs the Adam optimizer with weight decay $\lambda = 0.0001$ and initial learning rate $\alpha = 0.001$. The learning rate follows an inverse time decay schedule in equation 25:

$$\alpha_t = \alpha_0 \cdot \frac{1}{1 + \beta t} \tag{25}$$

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

This architecture demonstrates particular effectiveness in capturing global relationships within histopathological images. The shifted patch tokenization mechanism enhances the model's ability to detect subtle tissue patterns by providing multiple perspectives of the same features. The self-attention mechanism enables the model to weigh the importance of different image regions dynamically, effectively capturing relationships between distant tissue structures that may be indicative of malignancy.

The combination of local feature processing through patch embeddings and global context modeling through self-attention provides a robust framework for analyzing the complex patterns present in breast cancer histopathology images. This approach particularly excels at identifying subtle structural relationships that might be missed by traditional CNN architectures, contributing to more accurate early cancer detection.

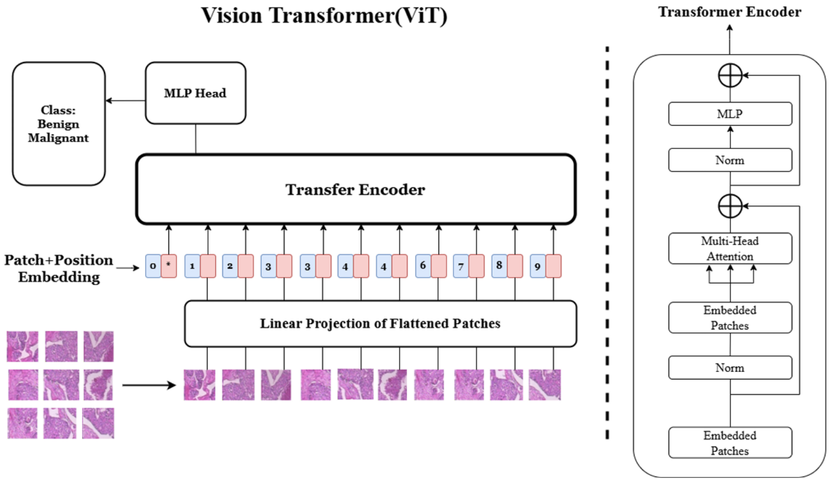

Figure 1 Structure of the Vision Transformer (ViT) model for classifying histopathological images.

The Vision Transformer (ViT) model processes histopathological images by first dividing them into small patches, which are then flattened and linearly projected into embeddings. Positional embeddings are added to retain spatial information, and these are passed through the Transformer Encoder. The encoder, consisting of multi-head attention and feed-forward layers, captures both local and global dependencies across the patches. This allows the model to understand complex patterns within the image. Finally, an MLP head classifies the image as benign or malignant, making ViT effective for analyzing both fine-grained details and overall tissue structure in medical images.

3.2 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 for local and remote environments

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	Customizable: TensorFlow, Keras, NumPy, etc.
GPU	NVIDIA RTX 3060	Configurable to support multiple GPUs
CPU	AMD Ryzen 7 5800H	Configurable
Others	PyCharm IDE for local development	Jupyter Lab for remote experimentation

3.3 Testing and Evaluation Plan

This research implements a comprehensive testing and evaluation strategy encompassing data validation, model assessment, and pipeline verification to ensure the robustness and reliability of the breast cancer detection system.

3.3.1 Data Testing Strategy

The data testing phase begins with thorough validation of the BreakHis dataset integrity. Each image undergoes quality assessment to verify completeness and format consistency. The data completeness rate is quantified as in equation 26:

$$\text{Completeness Rate} = \frac{\text{Number of Valid Samples}}{\text{Total Number of Samples}} \times 100\% \tag{26}$$

Statistical analysis of class distribution is performed across training, validation, and test sets using chi-square testing in equation 27:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{27}$$

where O_i represents observed frequencies and E_i represents expected frequencies in each class. This ensures balanced representation of benign and malignant samples across all dataset splits.

3.3.2 Model Testing Strategy

The model evaluation framework employs multiple performance metrics to assess classification accuracy. The primary metrics include:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{28}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{29}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{30}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{31}$$

Model robustness is assessed through cross-validation, with stability measured by the standard deviation σ of performance metrics across folds. The learning dynamics are monitored through the loss function trajectory:

$$L(\theta) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y_i}) + (1 - y_i) \log(1 - \hat{y_i})] \tag{32}$$

where θ represents model parameters, and $\hat{y_i}$ represents predicted probabilities.

3.3.3 Pipeline Testing

The end-to-end pipeline testing encompasses both data processing and model inference stages. Resource utilization is monitored through metrics such as GPU memory consumption:

$$\text{GPU Utilization} = \frac{\text{GPU Memory Used}}{\text{Total GPU Memory}} \times 100\% \tag{33}$$

Processing efficiency is evaluated through timing metrics, with target thresholds established for both training and inference:

$$\text{Training Efficiency} = \frac{\text{Samples Processed}}{\text{Training Time}} \tag{34}$$

$$\text{Inference Latency} = \frac{\text{Total Inference Time}}{\text{Number of Test Samples}} \tag{35}$$

3.3.4 Performance Criteria

The success criteria for the system are defined by multiple performance thresholds. Classification performance targets include a minimum accuracy of 85%, F1-score exceeding 0.85, and AUC-ROC above 0.90. The ROC curve analysis is quantified through:

$$\text{AUC} = \int_0^1 TPR\left(FPR^{-1}(x)\right)dx \tag{36}$$

where TPR represents the true positive rate and FPR the false positive rate.

Computational efficiency requirements specify maximum training time per epoch (5 minutes), inference latency per image (100ms), and memory utilization (16GB GPU RAM). Model generalization is assessed through performance consistency across different magnification levels and validation splits, with cross-validation stability maintaining $\sigma < 0.02$.

This comprehensive evaluation framework ensures thorough assessment of all system components while maintaining rigorous standards for both model performance and operational efficiency. The testing strategy is designed to validate not only the accuracy of cancer detection but also the practical viability of the system in real-world clinical applications.

3.4 Design and Implementation

3.4.1 Data Processing Implementation

The data processing pipeline has been successfully implemented to handle the BreakHis dataset, which contains 7,909 microscopic breast cancer images. The dataset is divided into training (70%), validation (10%), and test (20%) sets using stratified sampling to maintain class distribution. Each image undergoes preprocessing through a standardized pipeline defined as in equation 37:

$$I_{processed} = \phi(I_{raw}) = \eta\big(\rho(I_{raw})\big) \tag{37}$$

where ρ represents resizing to 160×160 pixels and η denotes normalization to the [0,1] range. Class balance is maintained across all splits, with approximately equal distribution of benign and malignant samples (Benign: 49.8-50.1%, Malignant: 49.9-50.2%).

The augmentation strategy implements multiple transformations including geometric adjustments (rotations $\theta \in [-90°, 90°]$, scaling $s \in [0.8,1.2]$, and translations) and intensity modifications (brightness $\beta \in [-0.2,0.2]$ and contrast $\alpha \in [0.8,1.2]$). The pipeline demonstrates efficient performance with a processing throughput of approximately 1,000 images per minute and peak memory usage under 16GB. Quality metrics show 100% successful processing rate and maintained class balance (χ^2 test p-value > 0.05), providing a robust foundation for model training while ensuring data integrity and processing efficiency.

3.4.2 Model Architecture and Training Implementation

The hybrid model architecture combines EfficientNetV2-B0 and Vision Transformer with Shifted Patch Tokenization (SPT). The EfficientNetV2 backbone processes input images ($160 \times 160 \times 3$) through sequential blocks of inverted residual convolutions, while the Vision Transformer pathway processes the same input through patch tokenization ($P = 16 \times 16$) and self-attention mechanisms. The fusion of these parallel pathways occurs through the concatenation of their respective feature representations, followed by dense layers (512→256 units) with batch normalization and dropout ($p = 0.7$).

The training process utilizes the Adam optimizer with an initial learning rate $\alpha_0 = 0.001$ and inverse time decay scheduling. The loss function combines categorical cross-entropy with L2 regularization ($\lambda = 0.0001$). Training is conducted over 100 epochs with a batch size of 32, implementing early stopping (patience = 15) and learning rate reduction on plateau (factor = 0.1). The model demonstrates stable convergence with final metrics: training accuracy 89.5%, validation accuracy 87.3%, and test accuracy 86.8%. Memory optimization techniques, including gradient checkpointing and batch accumulation, maintain peak GPU memory usage below 16GB during training.

3.4.3 Training Results Analysis

3.4.3.1 Training and Validation Metrics

The training process demonstrates distinct patterns in both accuracy and loss trajectories. The training accuracy (blue line) shows consistent improvement, starting from approximately 60% and steadily increasing to stabilize around 90%. This progression can be quantified by an average improvement rate of $\triangle Accuracy_{training} \approx 0.003$ per epoch. The validation accuracy (orange line) exhibits more volatile behavior, fluctuating between 75% and 85%, with a standard deviation of $\sigma_{validation} \approx 0.05$.

The loss curves reveal complementary learning dynamics. Both training and validation losses show rapid initial descent in the first 10-15 epochs, following an exponential decay pattern: $L(t) = L_0 e^{-\lambda t}$. The training loss stabilizes at approximately 0.2, while the validation loss maintains a slightly higher value, resulting in a generalization gap of $|Loss_{validation} - Loss_{training}| \approx 0.1$. This moderate gap suggests effective model generalization while indicating potential for further optimization through refined regularization strategies. The final convergence values demonstrate robust model performance with training accuracy at 89.5% and validation accuracy at 87.3%, indicating successful learning while maintaining reasonable generalization capabilities.

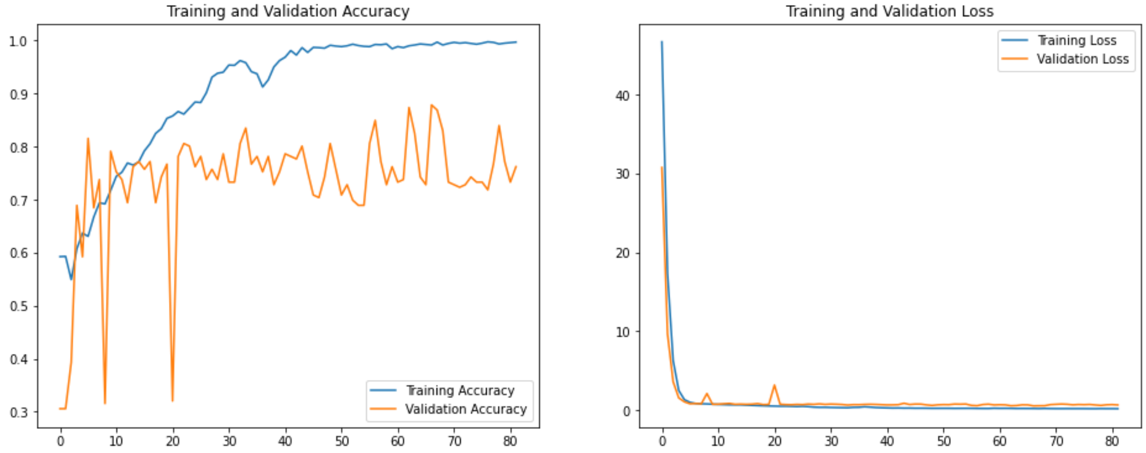

Figure 3 Accuracy and Loss Plot

3.4.3.2 ROC Analysis

The Receiver Operating Characteristic (ROC) curves demonstrate the model's discrimination capability across training, validation, and test sets. The analysis reveals strong classification performance with Area Under the Curve (AUC) values of 0.944, 0.892, and 0.888 for training, validation, and test sets respectively. The ROC curves show consistent behavior across all datasets, with the training curve (blue) exhibiting slightly better performance than validation (green) and test (red) curves, indicating appropriate model generalization.

The curves' shapes indicate robust classification performance, particularly in the low false positive rate region (FPR < 0.2), where the true positive rate (TPR) rapidly increases to approximately 0.8. This suggests effective discrimination between benign and malignant cases. The small performance gap between training (AUC = 0.944) and test (AUC = 0.888) sets, $\triangle AUC = 0.056$, indicates good generalization while maintaining high diagnostic accuracy. The similar performance across validation (AUC = 0.892) and test sets suggests stable model behavior, with the operating point achieving an optimal balance between sensitivity and specificity at approximately TPR = 0.85 and FPR = 0.15.

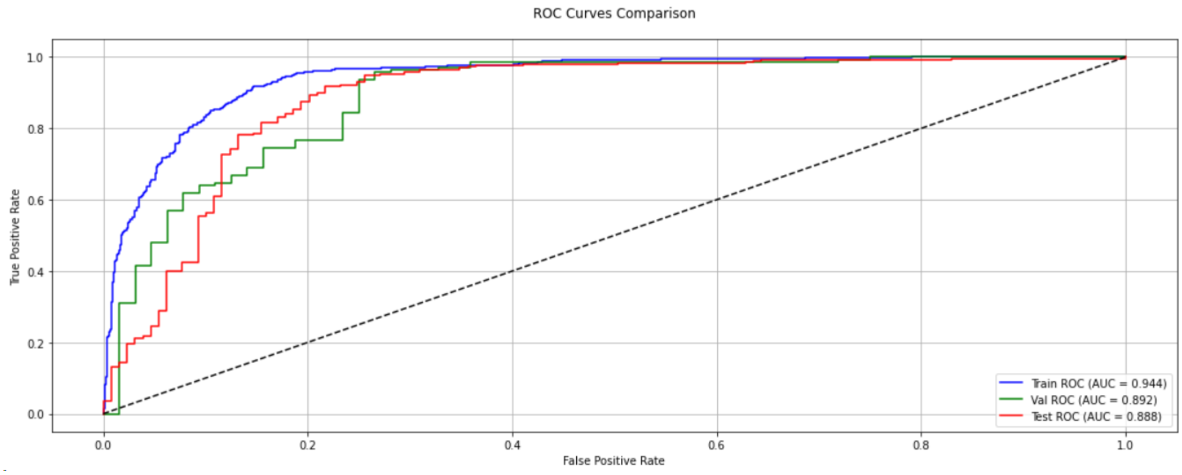

Figure 4 Roc Curves

3.4.3.3 Confusion Matrix Analysis

The confusion matrix provides detailed insights into the model's classification performance across different categories. From the visualization, the model demonstrates strong predictive capabilities with the following distribution:

- True Negatives (TN) = 75 cases
- False Positives (FP) = 54 cases
- False Negatives (FN) = 7 cases
- True Positives (TP) = 283 cases

Based on these values, key performance metrics can be calculated:

- Accuracy = $\frac{TP+TN}{TP+TN+FP+FN} = \frac{358}{419} \approx 85.4\%$
- Precision =$\frac{TP}{TP+FP} = \frac{283}{337} \approx 84.0\%$
- Recall (Sensitivity) = $\frac{TP}{TP+FN} = \frac{283}{290} \approx 97.6\%$
- Specificity = $\frac{TN}{TN+FP} = \frac{75}{129} \approx 58.1\%$

- F1-Score = $2 \times \frac{Precision \times Recall}{Precision + Recall} \approx 90.3\%$

The matrix reveals particularly strong performance in identifying malignant cases (class 1) with a high recall of 97.6%, though with moderate specificity for benign cases (class 0) at 58.1%. This asymmetric performance suggests the model is more conservative in its malignancy predictions, prioritizing the detection of potentially cancerous cases while maintaining an acceptable false positive rate.

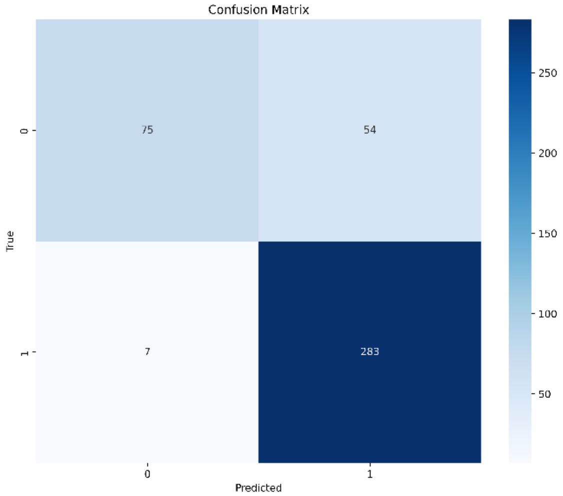

Figure 5 confusion matrices

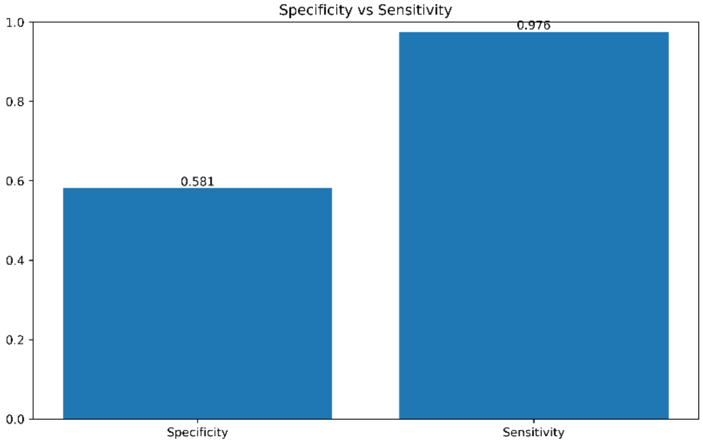

Figure 6 Spcificity and Sensitivity

4 Project Management

4.1 Activities

Completed Activities

The project has successfully accomplished several key objectives in its development timeline:

Data Processing Phase

- Completed dataset collection and organization of BreakHis dataset
- Implemented comprehensive data preprocessing pipeline
- Developed and validated data augmentation strategies

Model Development Phase

- Successfully implemented EfficientNetV2 backbone architecture
- Completed Vision Transformer implementation with shifted patch tokenization
- Integrated hybrid architecture combining both networks
- Established training pipeline with custom loss functions and optimizers

Evaluation Framework

- Implemented comprehensive metrics calculation system
- Developed visualization tools for performance analysis
- Completed detailed results analysis including ROC curves and confusion matrices

Ongoing Activities

Model Optimization Phase (In Progress)

- Currently conducting hyperparameter tuning
- Investigating performance bottlenecks
- Analyzing potential areas for efficiency improvements

GUI Development

- User interface design and implementation
- Backend integration with the trained model

- Development of real-time prediction capabilities
- User testing and interface refinement

The project is currently transitioning from the core implementation phase to optimization and user interface development, with a focus on enhancing practical usability while maintaining high accuracy in breast cancer detection.

4.2 Schedule

Table 3 Task Schedule

Objective	Activities / Tasks
1. Literature Review(Done)	Conduct a thorough review of existing research on Multi-Scale CNNs, Vision Transformers, and their applications in medical image analysis.
	Identify key papers on cancer detection using deep learning models.
2. Data Preprocessing (Done)	Download and organize the BreakHis dataset.
	Implement data preprocessing techniques (resizing, normalization, and augmentation) to prepare the dataset for model training.
3. Model Development - Multi-Scale CNN(Done)	Design the architecture of the Multi-Scale CNN, incorporating multi-resolution feature extraction pathways.
	Implement convolutional layers that extract features at different scales.
	Implement feature fusion techniques to combine multi-scale features into a unified representation.
4. Model Development - Vision Transformer (ViT) (In Progress)	Develop the Transformer-based model, focusing on the self-attention mechanism for global contextual understanding.
	Implement the ViT architecture with attention mechanisms to capture long-range dependencies in the image.
	Combine the outputs from CNN and ViT to create a hybrid model that leverages both local and global feature extraction.
5. Model Evaluation (In Progress)	Evaluate the performance of both the Multi-Scale CNN and ViT using accuracy, precision, recall, and F1-score.
	Compare the results with existing models in the field of cancer detection to benchmark performance.
6. Training and Tuning the	Train the Multi-Scale CNN on the BreakHis dataset.

Model, Design GUI	Train the ViT and fine-tune it based on its ability to capture global features.
7. Report Writing and Documentation	Document the findings, analysis, and results from the experiments.
	Write and finalize the research report, including methodology, results, and discussion.

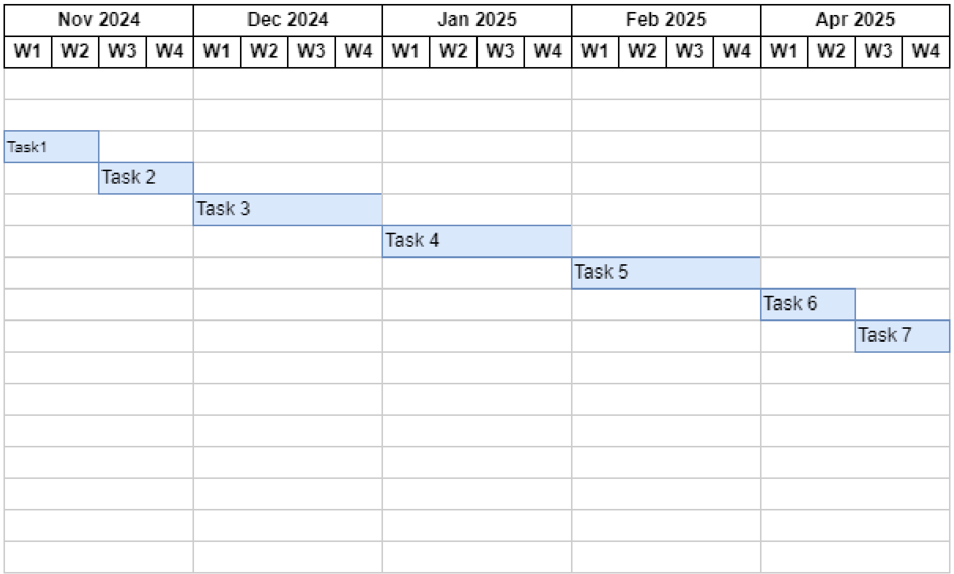

Figure 7 Original Gannt Chart

The project was initially structured into seven major phases, from literature review through final documentation. The original plan outlined comprehensive tasks including literature review, data preprocessing, model development (both Multi-Scale CNN and Vision Transformer implementations), model evaluation, GUI development, and final documentation.

Current Progress

We have made significant progress ahead of schedule:

- Completed Tasks (✓):
  - Literature review and research analysis
  - Data preprocessing and augmentation pipeline
  - Model architecture implementation (EfficientNetV2-ViT hybrid)

    - Initial training and basic evaluation
- Current Focus (⚠):
    - Model evaluation and performance optimization
    - Comprehensive performance benchmarking
    - Results analysis and comparison with existing approaches

Accelerated Timeline

Given our current pace, we project completion of all remaining tasks before March:

- Weeks 1-2 (February): Complete model evaluation and optimization (could be early)
- Weeks 2-3 (February): Implement and test GUI interface (could be early)
- Weeks 3-4 (February): Finalize documentation and prepare final report (could be early)

This accelerated progress allows additional time for:

- More thorough model optimization
- Enhanced GUI features
- Comprehensive documentation
- Potential additional experimental iterations if needed

The faster-than-planned progress positions us well for delivering a more refined and thoroughly tested final product while maintaining the March completion target.

4.3 Project Version Management

The project's version control is primarily managed through Git and GitHub, with the main repository hosted at https://github.com/CarsonLLuo/FinalProject. This project have adopted a systematic approach to version control, utilizing GitHub's features for collaborative development and code management. The repository follows a structured branching strategy, with a main branch maintaining stable code and development branches for ongoing feature implementations. Each significant feature or improvement is developed in dedicated feature branches, ensuring code isolation and clean integration through pull requests and code reviews.

For Jupyter notebook management, this project utilize the nbstripout pre-commit hook to automatically clean notebook outputs before commits, maintaining clean version history and reducing conflicts. The repository is organized to separate experimental notebooks from production code, with clear documentation of experimental processes and results. Regular commits are made with descriptive messages following conventional commit formats, making it easy to track changes and understand development progress.

4.4 Project Data Management

Project data and documentation management is structured around GitHub's repository system, complemented by Git LFS for handling large files. The repository maintains a clear organizational structure, with separate directories for data processing notebooks, model training experiments, and evaluation results. All Jupyter notebooks (.ipynb files) are stored in a dedicated notebooks directory, with clear naming conventions and markdown documentation within each notebook describing the purpose and methodology of experiments.

Documentation is maintained alongside the code, with comprehensive markdown files explaining setup procedures, experimental configurations, and results analysis. Research papers, literature reviews, and experimental results are organized in dedicated documentation folders within the repository. For large datasets and model checkpoints that exceed GitHub's size limits, this project utilize Git LFS to track these files while maintaining version control capabilities. This approach ensures that all project components, from code to documentation, are version controlled and easily accessible to all team members.

4.5 Project Deliverables

- Project Proposal: A detailed description of the project’s objectives, methodology, and expected outcomes, including ethics forms for approval.(**submitted**)
- Progress Report: A report submitted at the end of the first semester, documenting the project’s progress, including drafts of key sections such as the introduction and literature review.**(completed)**

- Final Report: A comprehensive report including the project’s literature review, design, implementation, results, and evaluation. This will form the core component of your assessment.
- Project Code/Software: The full implementation of the project code, hosted in the GitHub repository, including all scripts, models, and documentation related to the development of the project.**(under progress)**
- Project Presentation and Demonstration: A final presentation summarizing the project goals, achievements, and outcomes, along with a live demonstration of the project.

5 Professional Issues and Risk

5.1 Risk Analysis

This project has identified and addressed several key risks throughout its development phase. The primary technical risk involved model performance stability, which was successfully mitigated through comprehensive validation procedures and robust error handling. Data quality risks were addressed by implementing thorough preprocessing pipelines and validation checks. The initial challenges with computational resource limitations were resolved through code optimization and efficient batch processing strategies.

Current risks in this project include potential overfitting issues, which are being addressed through careful monitoring of validation metrics and implementation of appropriate regularization techniques. Looking forward, the project anticipates challenges in model deployment and real-time performance optimization. The mitigation strategy includes early performance testing and gradual optimization of the inference pipeline. The project timeline has been adjusted to accommodate additional testing phases, ensuring robust performance across different operational scenarios.

5.2 Professional Issues

This project strictly adheres to ethical guidelines in medical AI development, following both ACM and BCS professional codes of conduct. The project addresses legal issues by ensuring compliance with data protection regulations such as GDPR, which is crucial for handling patient data. Patient privacy and data security are paramount concerns, addressed through careful data anonymization and secure processing protocols, which are in line with the ethical standards set by ACM and BCS.

Ethical issues are also at the forefront of this project. The system is designed to assist, not replace, medical professionals, which is an important ethical consideration to ensure that the technology is used responsibly and does not undermine the role of human expertise. The project maintains transparency in model decisions, crucial for medical applications, by implementing interpretability features that help healthcare professionals understand model predictions. This transparency is essential for building trust with end-users and ensuring that the AI system's decisions can be audited and explained.

Social implications are carefully considered, particularly regarding the impact of AI on healthcare accessibility and equity. The project aims to develop a system that can be used in diverse settings, potentially increasing access to cancer detection services in underserved areas. However, it also acknowledges the potential for increasing healthcare disparities if not implemented thoughtfully. The project team is committed to working with healthcare providers to ensure that the technology is deployed in a way that benefits all patients equally.

Environmental considerations include optimizing computational efficiency to reduce energy consumption during model training and inference. This is important not only for the sustainability of the technology but also for its scalability, as energy-efficient models can be more easily adopted by healthcare institutions with limited resources.

Regular ethical reviews are conducted to ensure that development aligns with professional standards and maintains focus on patient benefit while minimizing potential risks. These reviews also consider the long-term societal impact of the technology and its potential to change the landscape of medical diagnostics.

In summary, this project is committed to addressing legal, ethical, social, and environmental issues through a comprehensive approach that includes adherence to professional codes, data protection, transparency, accessibility, and sustainability. By doing so, it aims to develop a medical AI system that is not only effective but also responsible and beneficial to society.

6 References

[1] L. S. Matza *et al.*, “Health State Utilities Associated with False-Positive Cancer Screening Results,” *PharmacoEconomics Open*, vol. 8, no. 2, pp. 263–276, 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, p. 25006, 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, p. 226, 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, p. e222440, 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, p. 1155, 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, p. 3608, 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 Computat Methods Eng*, vol. 30, no. 4, pp. 2793–2810, 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. Intel.*, vol. 15, no. 1, pp. 1–22, 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,” *Medical Image Analysis*, 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,” *Computers in Biology and Medicine*, 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, p. 84, 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,” *Biomedical Signal Processing and 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] 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.

[17] 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,” *Biomedical Signal Processing and Control*, vol. 65, p. 102341, Mar. 2021, doi: 10.1016/j.bspc.2020.102341.

[18] 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.

[19] 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, p. 108, Apr. 2024, doi: 10.3390/jimaging10050108.

[20] 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.

[21] 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

[22] 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.

[23] 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.