DEEP LEARNING FOR HISTOPATHOLOGICAL IMAGE CANCER DETECTION

GAN ZI HUI

UNIVERSITI TUNKU ABDUL RAHMAN

DEEP LEARNING FOR HISTOPATHOLOGICAL IMAGE CANCER DETECTION

GAN ZI HUI

A project report submitted in partial fulfilment of the requirements for the award of Bachelor of Biomedical Engineering with Honours

Lee Kong Chian Faculty of Engineering and Science
Universiti Tunku Abdul Rahman

May 2025

DECLARATION

I hereby declare that this project report is based on my original work except for citations and quotations which have been duly acknowledged. I also declare that it has not been previously and concurrently submitted for any other degree or award at UTAR or other institutions.

Name : GAN ZI HUI

ID No. : 21UEB01779

Date : 12 May 2025

COPYRIGHT STATEMENT

This final year project report is submitted in partial fulfilment of the requirements for the degree of Biomedical Engineering at Universiti Tunku Abdul Rahman (UTAR). This final year project report represents the work of the author, except where due acknowledgement has been made in the text. No part of this final year project report may be reproduced, stored, or transmitted in any form or by any means, whether electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the author or UTAR, in accordance with UTAR's Intellectual Property Policy.

ACKNOWLEDGEMENTS

I would like to express my gratitude to Associate Professor Ir. Dr. Tee Yee Kai, my research supervisor, for his invaluable advice, guidance and enormous patience throughout the development of the research.

I would also deeply grateful the faculty and the departmental members from Lee Kong Chian Faculty of Engineering and Science and Department of Mechatronics and BioMedical Engineering, for creating a pleasant working environment throughout my years in UTAR.

Furthermore, my sincere thanks go to Ir. Dr. Goh Choon Hian and the members at MIMOS for their generous assistance and support throughout this research.

Last but not least, I would like to thank my family and friends for their continuous encouragement and support, which played a pivotal role in the success of this project.

ABSTRACT

Head and neck cancers (HNC) are among the most prevalent cancers globally, with high mortality and poor prognosis often resulting from late-stage diagnoses. However, diagnostic difficulties are compounded by the histological complexity of HNCs and the subjective nature of manual histopathological analysis, which is prone to human error and inter-observer variability. Therefore, this study proposed a deep learning approach to assist in the classification of HNC from histopathological whole slide images, aiming to improve diagnostic accuracy and reduce observer bias. This study adopted the Head and Neck Squamous Cell Carcinoma dataset from the Clinical Proteomic Tumor Analysis Consortium, which consists of 390 whole slide images from various head and neck cancer sites, including 122 benign and 268 tumor slides. Convolutional neural network (CNN) models were trained using a transfer learning strategy, incorporating variants from the DenseNet, EfficientNet, MobileNet, ResNet, and VGG families. These models were fine-tuned using pre-trained weights and further evaluated for classification performance at three magnification levels (1.25×, 2.5×, and 5×). The top-performing CNN models were then combined using ensemble learning techniques to improve overall accuracy and robustness. The ensemble approach, particularly the majority voting with five models ensemble, outperformed individual models, achieving an accuracy of 96.09%, along with improved performance in sensitivity, precision, and F1-score. Visual interpretability tools, such as Gradient-weighted Class Activation Mapping, were employed to provide insights into the models' decision-making processes, enhancing the transparency and trustworthiness of the artificial intelligence predictions. The study also compared the CNN-based models to Vision Transformer models, showing that CNN ensembles achieved superior performance in classification tasks. This research highlights the potential of deep learning, particularly ensemble methods, in histopathological image analysis, with significant applications in computer-aided diagnosis for cancer detection. Further work should focus on addressing class imbalance, integrating the models into a clinical pipeline, and exploring multimodal learning to enhance model performance and clinical applicability.

Keywords: deep learning, convolutional neural networks, ensemble learning, whole slide image, head and neck cancer,

Subject Area: Q300-390 Cybernetics

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS / ABBREVIATIONS

AUC ROC	area under the receiver operating characteristic curve
FN	false nagative
FP	false positive
TN	true negative
TP	true positive
Acc	accuracy
DOR	diagnostic odds ratio
MCC	Matthew's correlation coefficient
Pr	precision
Re	recall
Sp	specification
λ	regularization rate
AI	artificial intelligence
CAD	computer-aided dianogsis
CNN	convolutional neural network
CPTAC	Clinical Proteomic Tumor Analysis Consortium
DL	deep learning
Grad-CAM	gradient weighted class activation mapping
HNC	head and neck cancer
HNSCC	head and neck squamous cell carcinoma
ML	machine learning
MV	majority voting
OSCC	oral squamous cell carcinoma
PyHIST	Python histological image segmentation tool
ReLU	rectified linear unit
r-HNC	rare head and neck cancer
TCIA	The Cancer Imaging Archive
UA	unweighted averaging
ViT	vision transform
WA	weighted averaging

WSI	whole slide image

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 General Introduction

Head and neck cancers, also known as Head and Neck Squamous Cell Carcinoma (HNSCC), are a group of cancers that typically begin in the squamous cells lining the moist surfaces of regions like the mouth, throat, and voice box (National Cancer Institute, 2021). These cancers represent the sixth most commonly diagnosed globally, with approximately 900,000 new cases and over 400,000 deaths reported annually (Stenson, 2025). The primary risk factor for HNSCC is tobacco use, accounting for approximately 75% of all cases, while other common risk factors include alcohol consumption and infections such as human papillomavirus and Epstein-Barr virus (Barsouk *et al.*, 2023). The survival rates for head and neck cancers vary significantly depending on the stage at diagnosis. Studies show that patients diagnosed with localized disease have a 5-year survival rate of 86.3%, which decreases to 69.0% for those with locally advanced disease and drops further to 40.4% for metastatic cases. (Barsouk *et al.*, 2023). Moreover, the diagnosis of head and neck cancers (HNCs), especially rare subtypes (r-HNCs), is also challenging due to their low incidence, overlapping histological features with more common cancers, and complex molecular profiles (Filippini *et al.*, 2024). The rarity of these tumors often results in limited clinical experience, leading to difficulties in accurate recognition and classification. Therefore, early detection and accurate dianogsis are important for improving prognosis and tailoring appropriate treatment strategies, especially given the significant survival differences across disease stages.

Currently, cancer diagnosis primarily depends on imaging techniques and pathological assessments. Modern cancer diagnosis involve a series of steps designed to detect and confirm the presence of cancer. It typically starts with a clinical examination, where a doctor evaluates the patient's symptoms and medical records. Imaging screening, such as computed tomography (CT) scans and magnetic resonance imaging (MRI) are then used to detect and evaluate suspected cancers. Moreover, a biopsy is usually required to confirm a cancer

diagnosis, as it provides a sample of abnormal tissue for further evaluation. The tissue is then analyzed through histopathological image analysis, where a pathologist examines it under a microscope to discover malignant cells and evaluate the specific type, level, and possible stage of the cancer (Robinson, 2024). In recent years, this process has been enhanced through the adoption of Whole Slide Imaging (WSI), which digitizes tissue slides at high resolution. WSI has seen major improvements in image quality and scanning speed, and studies have shown strong diagnostic agreement between WSI and traditional microscopy, supporting its growing use in clinical and research applications for cancer detection and classification (Rizzo *et al.*, 2022).

With advances in technology, particularly through artificial intelligence (AI) and machine learning (ML), cancer diagnostics are being revolutionized through the integration of complex algorithms and large datasets. Computer-aided diagnosis (CAD), an automated tool that uses computer-generated outputs, is gaining popularity due to its extensive use in digital image analysis across MRI, X-ray, endoscopy, ultrasound, and WSI to enhance clinical diagnosis (Halalli and Makandar, 2018; Komura and Ishikawa, 2018). Thereby improving early detection and diagnostic accuracy despite the computer's performance not surpassing that of experienced radiologists (Doi, 2007). Therefore, the integration of AI and ML into CAD systems is expected to further elevate their capabilities, enabling more accurate and timely diagnoses. As these technologies continue to evolve, they hold the promise of transforming cancer diagnostics by providing additional layers of analysis and reducing the reliance on traditional methods alone (Sebastian and Peter, 2022). The ongoing advancements in CAD are paving the way for more personalized and effective cancer detection strategies, ultimately contributing to better patient prognosis and advancing the discipline of oncology.

1.2 Importance of the Study

Since the AI revolution in the mid-20th century, machine learning has revealed its vast potential in the medical field, driving advancements in personalized treatment, predictive analytics, remote patient monitoring, and especially enhanced diagnostics (Boyle, 2024). Among these applications, deep learning (DL) in cancer detection has been extensively explored due to its ability to offer

advanced imaging interpretation by analyzing large datasets, including thousands of medical images and patient records. DL models can detect subtle patterns and nuances that may be overlooked by the human eye, making it a powerful tool in early cancer diagnosis. For instance, in automated histopathological analysis, DL algorithms can examine tissue samples at a microscopic level, identifying cellular abnormalities with remarkable precision. These systems can differentiate between benign and malignant cells, assess tumor aggressiveness, and even predict patient outcomes based on histological features (Ong *et al.*, 2023; McCaffrey *et al.*, 2024). This technology also benefits doctors by enabling earlier and more accurate cancer diagnoses, which allows for quicker treatment initiation before the disease spreads. Additionally, AI helps in reduce unnecessary follow-up biopsies by minimizing false positives, saving time, costs, and reducing patient anxiety(Spectrum AI, 2024). By speeding up the diagnostic process and prioritizing high-risk cases, AI-powered technologies improve the productivity of medical personnel by allowing them to focus on challenging situations, ultimately improving patient outcomes and streamlining healthcare systems (Alowais *et al.*, 2023).

These findings will not only contribute to the field of medical imaging and artificial intelligence but also support the broader goal of personalized medicine by enabling tailored diagnostic and treatment approaches based on individual patient profiles. AI technologies have significant potential to transform traditional cancer diagnostics, which have traditionally relied on the expertise of pathologists and are often labor-intensive and prone to human error. Nowadays, various machine learning approaches, including supervised and weakly supervised learning, unsupervised methods, transfer learning, and Vision Transformers, are actively being explored for their potential to support cancer diagnosis through tasks such as tissue segmentation, tumor classification, and feature extraction(Tiwari *et al.*, 2025). By leveraging DL techniques or other AI advancements, automated systems hold great potential for enhancing the early and accurate detection of cancer, ultimately improving patient outcomes and increasing survival rates.

1.3 Problem Statement

In Malaysia, oral squamous cell carcinoma (OSCC), the most common type of HNSCC, accounts for approximately 10.6% of cancer-related deaths in government hospitals, with a concerning 67.1% of cases diagnosed at an advanced stage (Ahmad *et al.*, 2021). The high rate of late-stage diagnoses significantly compromises patient survival and limits treatment effectiveness, highlighting the urgent need for improved early diagnostic strategies. Moreover, diagnostic challenges are particularly prominent in rare subtypes of HNSCC that originate from anatomically complex regions such as the nasopharynx, nasal and paranasal sinuses, salivary glands, and middle ear. These tumors often present overlapping histological features with more common cancers, making accurate identification difficult. Their rarity and anatomical complexity further complicate early detection, leading to diagnostic delays and difficulties in planning timely and effective treatment. Therefore, early and accurate identification of tumor origins, particularly in anatomically complex and histologically diverse regions such as the head and neck, is crucial for improving prognosis and guiding effective therapeutic decisions.

Furthermore, histopathology, a primary method for diagnosing cancer and determining its stage, can sometimes lead to misinterpretation and misdiagnosis, with a significant false-positive rate of approximately 27% (Wright, 2021). The reliance on manual interpretation of histopathological slides by pathologists further exacerbates this issue, as the process is time-consuming, subjective, and susceptible to inter-observer variability and human error (Wang *et al.*, 2025). Misdiagnoses can arise from the complexity of interpreting histopathological images, which may present subtle variations that are challenging to distinguish (Li *et al.*, 2023). For instance, histopathological images of early-stage cancers can resemble non-cancerous conditions or other diseases, complicating diagnosis. The subjective nature of image interpretation by pathologists, variability in expertise, and potential fatigue can lead to oversight or errors (Najjar, 2023). These misdiagnoses can delay appropriate treatment, allow cancer to progress, and impact patient outcomes significantly.

To solve these difficulties, deep learning has the potential to transform this field by enhancing early detection and precise diagnosis using advanced image analysis and pattern recognition, which may improve treatment accuracy

and patient outcomes. Therefore, there is a critical need for accurate and scalable deep learning models that can assist pathologists in detecting cancer with high precision and consistency.

1.4 Aim and Objectives

In this study, it focuses on the development of a CAD system based on deep learning for binary classification of histopathological images related to cancer, with a primary emphasis on HNSCC. State-of-the-art techniques such as transfer learning, ensemble learning, and Grad-CAM visualizations are employed to improve the accuracy and interpretability of the model. This method aims to contribute to advancements in automated cancer detection, improving both diagnostic precision and the ability to interpret model decisions in a clinical settings. Several objectives are aim to achieve in this study:

(i) To design and implement CNN and ViT models using transfer learning and fine-tuning for histopathological image classification.

(ii) To enhance classification performance through ensemble learning techniques.

(iii) To evaluate the performance of designed models.

(iv) To compare the performance of CNN-based approaches with ViT approaches on the same dataset.

1.5 Scope and Limitation of the Study

This study focuses on the development and evaluation of CNN models for the automated analysis of HNSCC histopathological images, with a specific emphasis on binary classification to differentiate between cancerous and non-cancerous tissues. It will also involve pre-processing WSIs to prepare them for input into the deep CNN model. Transfer learning techniques, especially fine-tuning, will be utilized to boost the model's learning capabilities. Top-performing models will then undergo ensemble learning to further enhance classification outcomes. The performance of these models will be evaluated and compared with a Vision Transformer (ViT) model to identify the most effective approach. While the primary focus is on HNC, the study has the potential to extend to other cancer types based on the results obtained.

However, the study encounters several limitations. The availability and quality of histopathological image datasets can impact the model's performance, as inadequate staining or poor-quality images may hinder effective training and validation. Additionally, the acquired dataset may show data imbalance, which could lead to biased model predictions and affect performance on underrepresented classes. To mitigate these issues, regularization techniques will be applied to reduce overfitting and improve the model's generalization across all classes. Moreover, this study will focus only on the HNSCC cohort, while the generalizability of the CNN model to other cancer types and histopathological conditions remains uncertain and requires further investigation. Lastly, while the model's performance has been evaluated within a research context, further validation and integration into clinical workflows are needed to ensure its reliability and practical effectiveness in real-world applications.

1.6 Contribution of the Study

This study contributes to the field of WSI analysis by exploring deep learning techniques for the binary classification of histopathological images, specifically focusing on HNSCC. The pipeline covers critical preprocessing steps, including tissue masking, tiling, patch extraction, stain normalization, and label assignment, followed by model training and evaluation. Several cutting-edge models are used, including CNNs (DenseNet, EfficientNet, MobileNet, ResNet and VGG) and ViT, which employ transfer learning and fine-tuning techniques. Ensemble learning techniques are applied to further improve classification performance by aggregating predictions from multiple CNNs. A comparative analysis between CNN-based and ViT-based approaches is conducted to assess their respective strengths in this domain. While the models generated in this study are not part of a fully deployed CAD system, the reliable models and reproducible pipeline developed here pave the way for future incorporation into diagnostic assistance systems. These models demonstrate significant potential for automating cancer detection, assisting pathologists with clinical processes, and contributing to the creation of interpretable, AI-powered decision support systems in digital pathology.

1.7 Outline of the Report

This report is structured as follows: Chapter 2 presents an overview of deep learning in histopathological image processing and analyzes related research, with a focus on CNN and ViT architectures, transfer learning strategies, and ensemble learning techniques. Chapter 3 outlines the deep learning pipeline for the study, including dataset preparation, model development, implementation procedures, and evaluation method. Chapter 4 presents the evaluation results, as well as the discussion and comparison of outcomes across architectures. Chapter 5 summarizes the key findings and outlines potential directions for future research. The confusion matrix, loss and accuracy curves, and ROC graphs are included in the Appendices for further reference.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Histopathological images are microscopic tissue samples used for disease analysis, commonly applied in cancer detection and stage determination. Traditionally, pathologists manually examine tissue samples under a microscope to identify abnormalities. Problems evolve including the complexity of these images, combined with the growing workload, makes the process time-consuming and may lead to findings influenced by the pathologist's subjectivity. To address these issues, the digitization of these slides has enabled the application of computational techniques to automate the analysis process, aiding pathologists in detecting abnormalities with greater accuracy. Whole slide imaging was first described by Wetzel and Gilbertson in 1999 as the digitization of entire histology slides or selected areas. The evolution of WSI has progressed from basic digitization of tissue slides to advanced systems with automatic refocusing, tissue recognition, and multimodal imaging, significantly improving efficiency and image quality in digital pathology. However, the complexity of histopathological images, combined with the need for precise interpretation, poses significant challenges, especially when dealing with large-scale datasets. Therefore, modern histopathological image analysis, driven by advancements in digital pathology and artificial intelligence, has become essential in clinical practice (Moscalu *et al.*, 2023).

In recent years, deep learning methods are potentially poised to revolutionize clinical practice by enhancing diagnostic accuracy, streamlining workflows, and improving patient outcomes, with ongoing research addressing challenges to integrate these advancements into routine medical settings (Moscalu *et al.*, 2023). A successful example by Guo et al. (2022) demonstrates the valuable clinical applications of CAD in MRI, where it is used to differentiate between noninvasive and invasive breast lesions, classify invasive cancers with or without lymph node metastasis, and assist in tumor staging. In result, the integration of digital pathology with AI enables pathologists to expand their diagnostic capabilities, facilitating the extraction of clinical

insights from large datasets, ultimately improving patient care and operational efficiency.

2.2 Deep Learning Architectures

Deep learning architecture refers to the structure and organization of layers in an artificial neural network designed to automatically learn patterns and representations from data, eliminating the need for manual feature extraction. (Madhavan and Jones, 2024). These architectures can generally be classified into supervised and unsupervised learning models. In cancer diagnostics, particularly using histopathological images, deep learning models like CNNs and ViTs have shown promising performance in image recognition and classification tasks. This section provides an overview of these prominent architectures used in the present study, outlining their design principles, variants, and applications in cancer detection.

2.2.1 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks have revolutionized medical image analysis by leveraging hierarchical feature extraction to enhance classification performance. Unlike traditional image processing methods, CNNs automatically learn and extract features from raw image data through multiple layers of convolutional operations, pooling, and activation functions (Kalra, 2023). In a CNN, the convolutional layer applies filters to the image to detect features, using a set of weights to perform mathematical convolutions and generate feature maps (Shajun Nisha and Nagoor Meeral, 2021). These maps are then processed by the ReLU activation layer, which introduces non-linearity by setting negative values to zero. The pooling layer subsequently diminishes the spatial dimensions of the feature maps, retaining only the most significant information and helping to prevent overfitting. After several convolutional and pooling operations, the network uses fully connected layers, where flattened features are classified using a softmax function to output probabilities for different classes. This capability allows them to identify complex patterns and structures within histopathological images, which are critical for accurate cancer diagnosis.

2.2.1.1 VGG

VGG is one of the earlier deep learning models that demonstrated the effectiveness of deep networks for image classification. It consists of sequential layers of convolutional filters that capture detailed features. While VGG is known for its simplicity and ease of implementation, its high computational complexity makes it less ideal for large histopathological datasets. The VGG models, developed by Simonyan and Zisserman (2014) and presented in their 2014 paper, aimed to explore how increasing the depth of convolutional neural networks impacts performance in large-scale image recognition tasks. The primary variants of VGG used for transfer learning are VGG-16 and VGG-19 (Shoeibi *et al.,* 2022). Both VGG-16 and VGG-19 are popular choices for transfer learning in histopathological image detection. For instance, Setiawan, Pramudita and Mulaab (2024) demonstrated the effectiveness of these models in automated lung cancer detection, achieving a highest accuracy of 97%. Similarly, Kanimozhi and Priyadarsini (2024) used VGG-19 for breast cancer detection, attaining an impressive accuracy of 99.22%. Therefore, VGG models, particularly VGG-16 and VGG-19, have proven effective for transfer learning in histopathological image detection. Despite their computational complexity, these models have achieved notable results in various cancer detection tasks, demonstrating their value in advancing automated diagnostic systems.

2.2.1.2 ResNet

ResNet (Residual Networks) was introduced by He et al. (2016) in their 2015 paper titled “Deep Residual Learning for Image Recognition.” It introduced the concept of residual learning, which allows for deeper networks without the vanishing gradient problem. Variants of ResNet include ResNet50, ResNet101, and ResNet152, each differing in the number of layers and the depth of the network. The different variants of ResNet primarily differ in their depth, or the number of layers, which affects their capacity and computational complexity (Chaure, 2024). These variants offer different trade-offs between computational complexity and model performance, allowing flexibility in various image recognition tasks. Findings indicate that ResNet-50 is frequently used for feature selection in the initial stage of multi-classification tasks. For instance, Shen et al. (2023) and Marostica et al. (2021) both employed ResNet-50 for

feature extraction before proceeding with further classification. This might due to ResNet's use of residual blocks allows it to train very deep networks effectively. This design addresses the vanishing gradient issue and allows the network to learn complex features more effectively compared to some other architectures. Moreover, Ashwini et al. (2024) demonstrated a framework using ResNet50 for early breast cancer detection, achieving an impressive accuracy of 96.9% with their proposed system. These capabilities underscore ResNet's significant role in advancing both image recognition and medical diagnostics.

2.2.1.3 DenseNet

DenseNet (Densely Connected Convolutional Networks) was proposed by Huang et al. (2017). It further improved upon previous architectures by introducing dense connections between layers, allowing for more efficient gradient flow and feature reuse. Its compact architecture makes it suitable for analyzing high-resolution histopathological images, reducing the need for large computational resources. For instance, Noaman et al. (2024) had shown their promising result in automated breast cancer detection via the fusion of DenseNet201 with color histogram techniques to chieve a 99.683% of accuracy. Besides, Potsangbam and Shuleenda Devi (2024) demonstrated the effectiveness of transfer learning with the DenseNet architecture, achieving an accuracy of 96.53% at 100x magnification. These examples illustrate DenseNet's effectiveness in enhancing diagnostic performance in specialized image analysis tasks.

2.2.1.4 MobileNet

MobileNet, proposed by Howard (2017), is a lightweight architecture designed for use in resource-constrained environments. Its depthwise separable convolutions reduce computational complexity while maintaining high accuracy, yielding it appropriate for real-time applications and deployment on mobile devices. The MobileNet variants include MobileNetV1, MobileNetV2, MobileNetV3, MobileNetV3-Large, and MobileNetV3-Small, each offering distinct features tailored to different needs and computational constraints. According to research by Datta Gupta et al. (2023), MobileNet achieved a high accuracy of 98%, matching that of InceptionV3 and outperforming ResNet50

by 13.34% in a three-class classification task, while having a model that is at least six times more compact than the others. These advantages highlight MobileNet's effectiveness in delivering high performance with significantly lower computational demands, making it particularly well-suited for mobile and edge computing applications.

2.2.1.5 EfficientNet

EfficientNet, as one of the latest CNN proposed by Tan (2019), scales both depth, width, and resolution of the network in a balanced manner, leading to state-of-the-art performance in various image classification tasks. Its efficiency in handling large image datasets with reduced parameters makes it particularly well-suited for histopathological image analysis, where computational resources may be limited. For instance, Albalawi et al. (2024) developed a deep learning model based on EfficientNetB3, achieving an impressive 99% accuracy in differentiating between normal epithelium and OSCC tissues using a substantial dataset of 1224 images from 230 patients. Moreover, Abhishek et al. (2024) demonstrated that the EfficientNetB4 model achieved superior performance with an accuracy of 99.89% in classifying colorectal cancer from histological images, outperforming other models such as GoogleNet, AlexNet, and various ResNet architectures, which all had accuracies below 95%. Therefore, EfficientNet's scalable architecture, with its variants from B0 to B7, allows for modifications in depth, width, and resolution, making it adaptable for various computational needs and accuracy requirements in histopathological image analysis.

2.2.2 Vision Transformers (ViTs)

Unlike CNNs, Vision Transformer is a deep learning architecture that adapts the Transformer model, originally designed for text, to image analysis by treating images as sequences of patch embeddings. Dosovitskiy et al. (2020) pioneered the ViT by introducing a pure Transformer architecture for image classification and showing that, when trained on massive datasets like ImageNet-21k, it can surpass CNN performance. ViT divides an image into fixed-size patches, flattens them, and feeds the resulting sequence into a standard Transformer encoder, which a simple fully connected neural network includes Multi-Head

Attention and Multi-Layer Perceptrons, to extract features for image classification (Shah, 2022). While ViT and CNNs share similar steps, such as splitting the image into patches, using pretrained models, and fine-tuning, they are often compared or even combined in hybrid models due to their complementary strengths.

ViT shows its prominent in cancer clascification. For instance, Abadi and Reza (2024) achieved 95.11% accuracy in breast cancer classification using ViT by employing a progressive fine-tuning strategy that gradually updated more layers to adapt to the cytological image domain. Besides, ViT has also demonstrated strong performance in multi-class classification, particularly in skin cancer detection, as evidenced by studies from Yang, Luo and Greer (2025) and Ozdemir and Pacal (2025) , which reported high classification accuracies of 95.05% and 93.48%, respectively. Several studies have also proposed hybrid CNN and ViT frameworks to leverage the advantages of both architectures for improved classification performance (Hayat *et al.*, 2024; Katar *et al.*, 2024; Patheda *et al.*, 2025). The study from Momentum (2022) also highlighted several limitations of ViT learning, including lack of inherent positional awareness, fixed input resolution constraints, loss of fine-grained spatial information, patch border disruption, lack of translational equivariance, and high computational cost due to quadratic scaling with input length, which pose a significant challenge to its scalability and effectiveness in high-resolution or dense prediction tasks.

2.3 Deep Learning Techniques

2.3.1 Transfer Learning

Transfer learning has become a critical approach in medical image analysis due to the limited availability of annotated medical datasets. This technique leverages deep learning models pre-trained on large, diverse datasets like ImageNet, which contain vast amounts of labeled data across various categories, and then applie it to another task. By using CNNs pretrained on large general-purpose image datasets, researchers can fine-tune the models on medical images, thereby significantly improving accuracy, reducing training time, and addressing the challenge of limited labeled data (Zheng *et al.*, 2023). For instance,Hava Muntean and Chowkkar (2022)demonstrated that transfer

learning with the DenseNet121 model achieved 86.6% accuracy in classifying breast histopathological images at the 100X magnification level, with a 16.4% increase in training accuracy compared to models trained from scratch. Besides, Deebani et al. (2025) analyzed the effectiveness of Transfer Learning and transformers in multiscale cancer detection, achieving 97.41% accuracy for colon cancer and 94.71% accuracy for histopathological lung cancer detection. These studies highlight the potential of transfer learning in reducing computational costs, minimizing the need for large datasets, and improving model generalizability. However, transfer learning can lead to negative transfer and reduced model performance if the source and target tasks are dissimilar, the data distributions differ significantly, or an inappropriate model is applied. Ongoing research is focusing on methods such as distant transfer and various evaluation techniques for assessing task and dataset similarities, with the goal of reducing negative transfer and increasing transfer learning efficacy.

2.3.2 Ensemble Learning

The intricate characteristics of histopathological images often make them difficult to identify recognizing features using a pre-trained classification method. Hence, ensemble learning approaches, which intergrate several classification models, are commonly employed to address the complexities natural in analyzing these images.

Ensemble learning, first introduced by Nilsson in 1965, is a supervised learning approach where multiple base models are trained and their predictions are combined to generate a more precise overall result(Yang, 2017). The fundamental idea is to leverage the combined strength of diverse models, each with unique error patterns, to achieve improved overall performance compared to individual models. In cancer detection, ensemble methods have shown to be effective by reducing model variance and improving robustness. Techniques such as majority voting, stacking, and weighted averaging are commonly used to integrate the outputs of different CNN models, resulting in a more reliable classification system. For instance, Yong et al. (2023) demonstrates that the ensemble models, comprising EfficientNetB0, EfficientNetB1, DenseNet121, DenseNet169, and MobileNet with unweighted averaging, significantly enhance gastric cancer detection accuracy from histopathological images,

achieving a state-of-the-art accuracy of 99.20% in 160 × 160 pixel patches and offering valuable support for early diagnosis. By using the accuracy from pre-trained models as weights for averaging in the ensemble, Zheng et al. (2023) achieved an impressive 98.90% accuracy with their deep ensemble model for binary classification of breast histopathological images. This model outperformed recent transformer and MLP models by 5%–20%, showcasing its superior performance in classification tasks.

Addictionally, three ensemble methods, including majority voting, averaging, and probability-based fusion, were employed to categorize cardiovascular tissues into six distinct classes in Mittal (2021) study. For majority voting, the final prediction was determined by selecting the most frequently predicted label from the constituent CNNs. To produce the final prediction, the probability-based fusion method normalized and combined the predicted probabilities from the CNNs. While the averaging ensemble of three CNNs achieved the highest overall F1-score, the method provided the best F1-score with six CNNs. Conversely, the majority voting method did not surpass the performance of the other two ensemble techniques in any configuration (Mittal, 2021). In short, ensemble learning methods, with their capacity to integrate multiple classification models, offer a robust solution for the complex task of analyzing histopathological images. By leveraging diverse models and combining their outputs through techniques like unweighted averaging, majority voting, and probability-based fusion, these methods enhance accuracy and reliability in cancer detection.

2.4 Summary

This literature review examined the development and application of deep learning techniques for histopathological image analysis, particularly in cancer detection. It reviewed major deep learning architectures, including CNNs and ViTs, outlining their structural differences, benefits, and prior success in medical image classification. The key techniques such as transfer learning and ensemble learning were also discussed, which enhance model generalization and performance, particularly on limited medical datasets. The summaries of each reviewed study are aligned with the specific cancer type targeted for

classification, as well as the selection of architectures and techniques adopted in the present study, as detailed in Table 2.1.

Table 2.1: Summary of Reviewed Models.

Reference	Cancer Type	Method	Accuracy
Setiawan, Pramudita, & Mulaab (2024)	Lung Cancer	VGG16, VGG19 for automated detection	97%
Kanimozhi & Priyadarsini (2024)	Breast Cancer	VGG19 for detection	99.22%
Ashwini et al. (2024)	Breast Cancer	ResNet50	96.9%
Noaman et al. (2024)	Breast Cancer	DenseNet201 + color histogram fusion for automated detection	99.683%
Potsangbam & Shuleenda Devi (2024)	Breast Cancer	DenseNet for transfer learning	96.53%
Datta Gupta et al. (2023)	Three-class Classification	MobileNet for classification task	98.00%
Albalawi et al. (2024)	Oral Squamous Cell Carcinoma	EfficientNetB3 for differentiation between normal and carcinoma tissues	99%
Abhishek et al. (2024)	Colorectal Cancer	EfficientNetB4 for histological image classification	99.886%
Abadi & Reza (2024)	Breast Cancer	ViT with progressive fine-tuning	95.11%

Reference	Cancer Type	Method	Accuracy
Yang, Luo & Greer (2025)	Skin Cancer	ViT for multi-class skin cancer classification	95.05%
Ozdemir & Pacal (2025)	Skin Cancer	ViT for multi-class skin cancer classification	93.48%
Hayat et al. (2024)	Breast cancer	EfficientNetV2L-ViT	99.83%
Katar et al. (2024)	Lung Cancer	EfficientNet-B0 + LBP + ViT Encoder + SVM	99.87%
Patheda et al. (2025)	Breast cancer	CNN+ViT-B16	90.1%
Hava Muntean and Chowkkar (2022)	Breast Cancer	Transfer Learning with DenseNet121	86.6%
Deebani et al. (2025)	Colon and Lung Cancer	Transfer Learning and Transformers (Multiscale)	Colon: 97.41% Lung: 94.71%
Yong et al. (2023)	Gastric Cancer	Ensemble Learning (EfficientNetB0, EfficientNetB1, DenseNet121, DenseNet169, MobileNet with unweighted averaging)	99.20%
Zheng et al. (2023)	Breast Cancer	Deep Ensemble Model (Unweighted Averaging)	98.90%

CHAPTER 3

METHODOLOGY AND WORK PLAN

3.1 Introduction

The methodology of this project focused on five key phases: image preprocessing, data labelling and spliting, transfer learning, ensemble learning, and model evaluation. The workflow began with the input of WSIs, which are subjected to image pre-processing techniques such as patch extraction and normalization to prepare them for model training. The processed image patches were then split into training, validation, and testing sets to ensure a robust evaluation process. Multiple pre-trained CNNs were fine-tuned through transfer learning to adapt them to the specific cancer detection task. The predictions from these individual models were then integrated using ensemble learning, where predictions from multiple CNNs were combined to form a more accurate and generalizable ensemble model. This ensemble generated the final model output, which was then assessed through model evaluation metrics to determine its classification performance. Overall, the proposed deep learning-based classification pipeline were designed to be robust, accurate, and scalable, aiming to assist pathologists in detecting cancer with high precision and reliability. The complete process flow is shown in Figure 3.1.

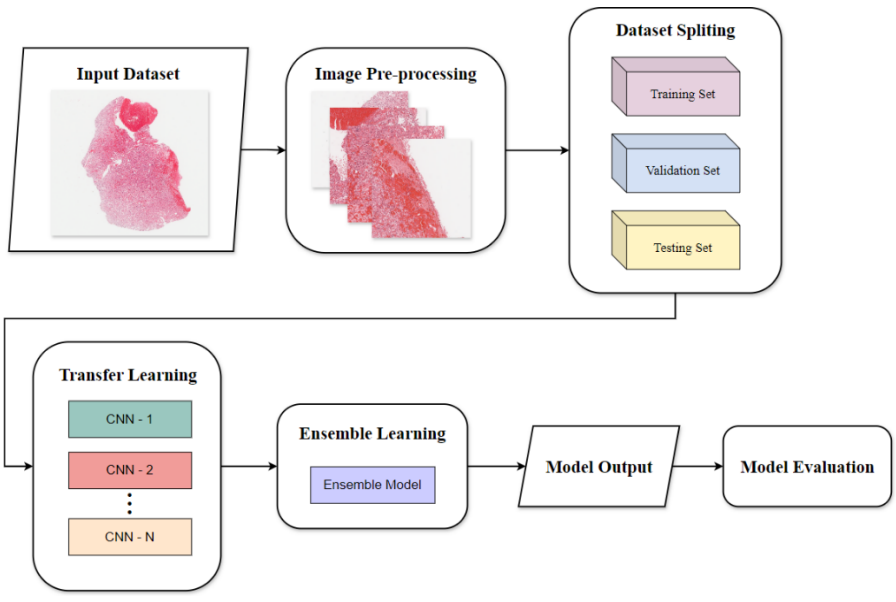


Figure 3.1: Process Flow.

3.2 Dataset Resource

The histology dataset utilized in this study was obtained from The Cancer Imaging Archive (TCIA) as part of the National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC) (2018) Head-and-Neck cancer (CPTAC-HNSCC) discovery study. It consists of 390 WSI, including 122 normal and 268 tumor slides in svs format, which are critical for our investigation. The CPTAC-HNSCC dataset includes tumor samples primarily from common head and neck cancer sites such as the oral cavity, tongue, buccal mucosa, oropharynx, floor of mouth, larynx, tonsil, alveolar ridge, and epiglottis. These samples encompass a variety of subtypes of HNSCC, including Keratinizing HNSCC, Acantholytic HNSCC, and Basaloid HNSCC, which vary in their histological characteristics and clinical behavior as shown in Figure 3.2. This diverse collection of images enhances the model's ability to learn distinct patterns and features characteristic of both normal and cancerous tissues.

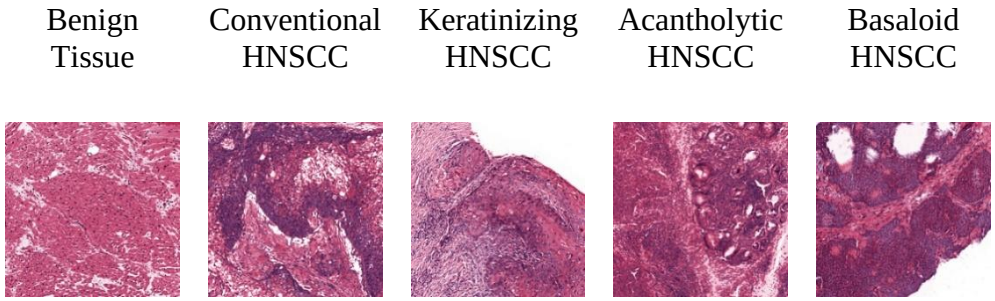


Figure 3.2: Example Slides for Each Subtype of Tissues in the Dataset.

3.3 Histophatological Image Pre-processing

The original WSI from CPTAC comprised high-resolution histopathological slides in the .svs format with a 20x magnification. Due to the high computational complexity associated with processing these large images, pre-processing was necessary to prepare the images for input into the transfer learning model. As shown in Figure 3.3, the image pre-processing in this study included four stages: tissue masking, tile extraction, segmentation, and stain normalization. Tissue masking, tile extraction, and segmentation were performed using PyHIST, which is a lightweight, semi-automatic command-line tool designed for extracting tiles from WSI in histopathology (Muñoz-Aguirre *et al.*, 2020). Stain normalization was performed with StainTools, a Python 3-based toolset for

tissue stain normalization and augmentation, which includes methods such as Macenko and Vahadane (Byfield, Godard and Gamper, 2021). After these pre-processing steps, the patches were labeled and split for training, validation, and testing purposes.

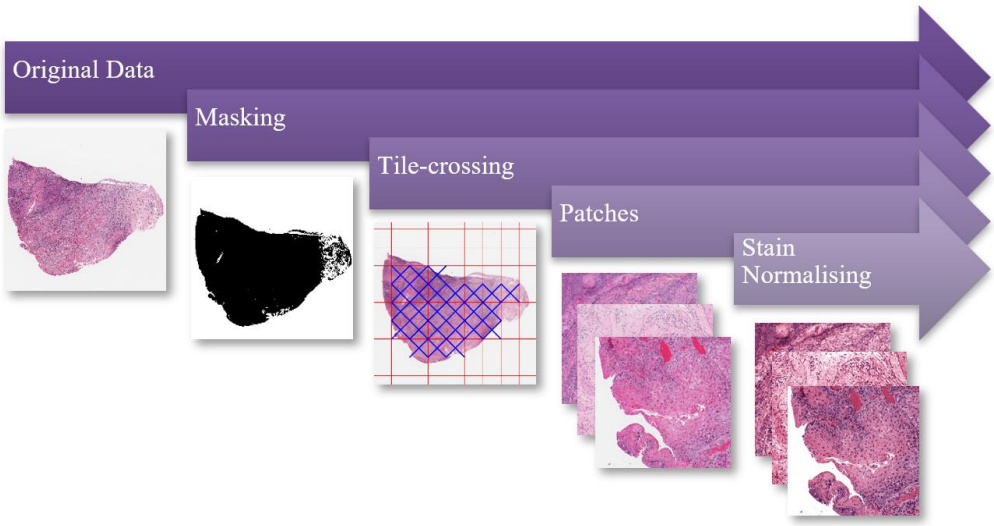


Figure 3.3: Image Pre-processing Steps.

3.3.1 Tissue Masking

Tissue masking is a critical preprocessing step in analyzing WSI in histopathology. It involves distinguishing tissue regions from non-tissue areas, such as the glass background, within a WSI. In this project, PyHIST was employed to generate tissue masks using a graph-based segmentation method. Additionally, Otsu's thresholding method was used to separate an image into foreground and background regions based on pixel intensity values. This automated algorithm works by maximizing the variance between the two classes (tissue and background) while minimizing the variance within each class. It determined the optimal threshold by analyzing the histogram of grayscale intensities and selecting the point that minimizes intra-class variance (Vijay and Patil, 2016). Pixels with intensities below this threshold were assigned to the background class, while those above are classified as foreground. This resulted in a clear distinction between tissue and non-tissue areas, facilitating the extraction of specific objects or regions of interest from the image (Gopalakrishnan, 2023). After generating the mask, it was transformed from the OpenCV format to the PIL format, and the corresponding background color array was stored subsequent tile extraction and processing tasks.

3.3.2 Tiles Extraction

A grid of 224×224 pixel tiles was created over the masked image. The magnification of the tiles could be adjusted by controlling the downsampling factors, with maximum magnification factor of 20× corresponding to a downsampling factor of 1. As shown in Figure 3.4, WSI were stored in a pyramidal structure that enables access to multiple resolution levels. The relationship between magnification and downsampling is illustrated Equation 1.1, where higher downsampling leads to lower image resolution. For this study, tiles were extracted at three different magnification levels, including 1.25×, 2.5×, and 5×, corresponding to downsampling factors of 16, 8, and 4, respectively. As illustrated in Figure 3.5, higher magnification tiles offered greater detail but also increase computational costs and processing time. Although this could enhance feature detection, excessively high magnification may not always improve results and could lead to diminishing returns due to increased noise and resource demands. Therefore, balancing magnification with computational efficiency was essential for optimal model performance. In this project, a downsampling factor of 8 was used to produce tiles with a 2.5x magnification.

$$Downsampling\ Factor = \frac{Original\ Magnification}{Target\ Magnification} \qquad (3.1)$$

With the tile grid and tissue mask in place, each tile was evaluated to ensure it met a minimum tissue coverage threshold, which was set at 0.5 in this project to ensure adequate tissue content. Tiles meeting this criterion were extracted from the whole-slide image at the desired resolution. While higher tissue coverage in patches was desirable for reducing non-informative areas, setting a higher threshold could reduce the number of usable tiles, potentially limiting the availability of histopathological images for model training, especially at lower magnification settings.

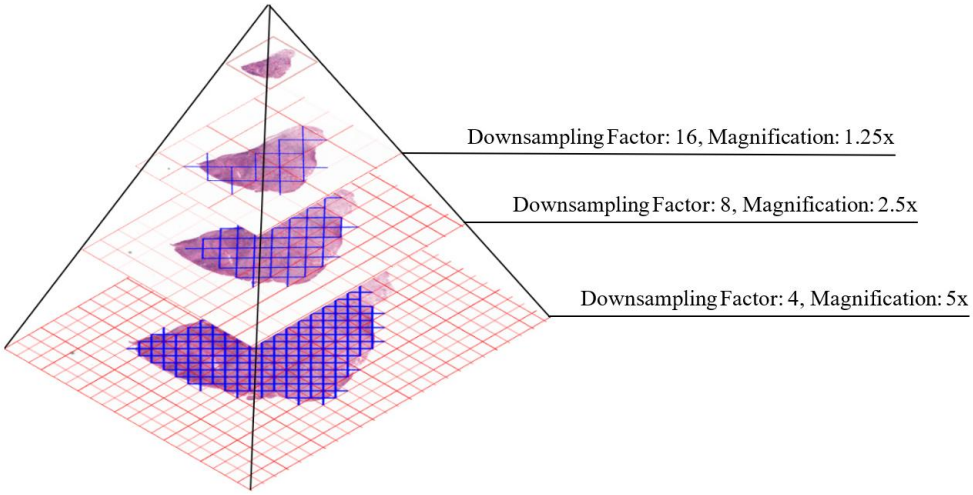


Figure 3.4: WSI Pyramidal Structure.

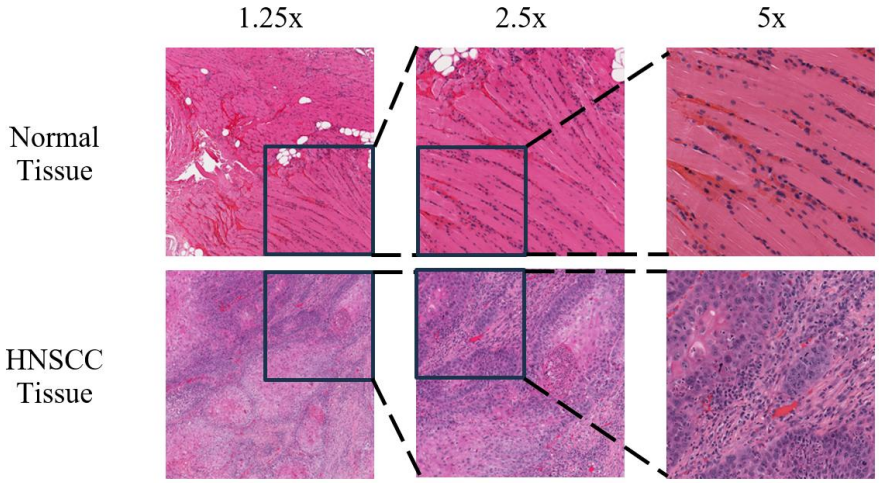


Figure 3.5: Image patches extracted at different magnification levels.

3.3.3 Patches Cleaning and Stain-Normalisation

A patch-cleaning step was conducted to eliminate undesired images that could negatively impact model performance. Specifically, tiles with excessive dark regions, commonly resulting from scanning artifacts such as blur, pen ink, or folded tissue, were identified and removed. This was achieved using a threshold-based approach, where any tile with over 90% of its pixels below a grayscale intensity of 50 was considered a "black image" and automatically excluded from the dataset. After automated filtering, the remaining patches were manually reviewed to ensure that any residual artifacts were also eliminated.

After cleaning, stain normalization was applied using the Vahadane method via the StainTools library. Vahadane's technique is built on sparse non-negative matrix factorization, which decomposes histological images into stain

color bases and their corresponding concentration maps (Vahadane *et al.*, 2016). By substituting the original stain color bases with those from a reference image, the method standardizes the color appearance across different samples while preserving the underlying tissue structure. In this study, a representative reference image was selected, and all patches were normalized to match its stain profile. Figure 3.6 illustrates the outcome of this process, showing the stain-normalised patches for each subtype of HNSCC. However, since stain weights were processed independently, the method might not fully consider the global intensity or relative proportion of each stain, occasionally resulting in overrepresentation of certain stains (Hoque *et al.*, 2024). Despite this limitation, Vahadane normalization remained effective for reducing inter-slide stain variability, allowing the model to focus on learning morphological patterns rather than being influenced by inconsistent staining.

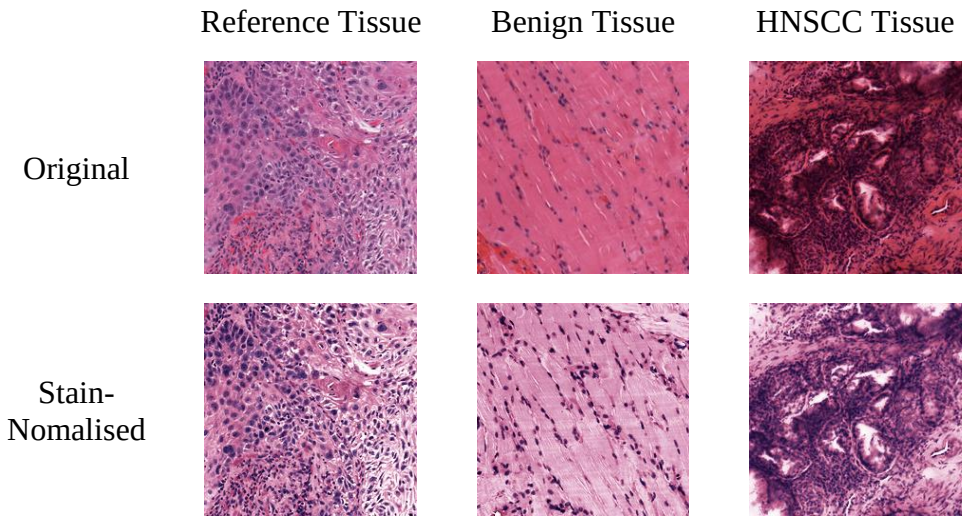


Figure 3.6: Output of Stain Normalisation for Each HNSCC Subtype.

3.3.4 Patch Labelling and Data Spliting

Once the patches were prepared, each was labeled according to its parent WSI using binary labels: "0" for normal tissue and "1" for malignant tissue. This labeling was applied to the entire WSI rather than to specific regions of interest (ROIs) within the slide. Although including the entire WSI may introduce some irrelevant data, as a tumor-labeled WSI might also contain normal tissue patches, studies from Phan et al. (2021) and Koo et al. (2023) had shown that it is not significantly affect the classification performance.

At the end of image preprocessing, the generated patches were divided into training, validation, and testing sets with a distribution ratio of approximately 70%, 20%, and 10%, respectively, and saved in an array for further processing. The distribution counts for each divided set are recorded in Table 3.1.

Table 3.1: Number of Patches in Training, Validation, and Testing Sets.

Dataset Splitting	WSI	Patches with 1.25x	Patches with 2.5x	Patches with 5x
Training	273	5331	21552	86520
Validation	82	1408	5949	22419
Testing	35	556	3273	15411
Total	390	7295	30774	124350

3.4 Classification Model Development

Transfer learning was carried out using five CNN architectures: DenseNet, EfficientNet, MobileNet, ResNet, and VGG, along with their respective variants. A shown in Figure 3.7, transfer learning was performed using models pretrained on the ImageNet dataset, with 80% of the convolutional base layers frozen and the remaining 20% unfrozen for fine-tuning on the histopathological dataset. Each pretrained model was selected and configured to accept input images of size 224 × 224 pixels. Four new trainable classification layers were then added, including a global average pooling layer, a dense connected layer with 512 neurons and ReLU activation, a final dense layer with 1 neuron and a sigmoid activation function for binary classification. The dense layer incorporated an L2 regularizer (with $\lambda = 0.01$) to penalize large weights and reduce overfitting.

Moreover, the models were compiled with binary cross-entropy as the loss function and AdaBound optimizer, with a learning rate of 0.00001. Training was conducted with a batch size of 32 over a maximum of 50 epochs, with early stopping applied if validation accuracy did not improve for 5 consecutive epochs. Table 3.3 summarized all the training parameters in the experiment. Model performance was evaluated on both validation and test datasets to ensure classification effectiveness.

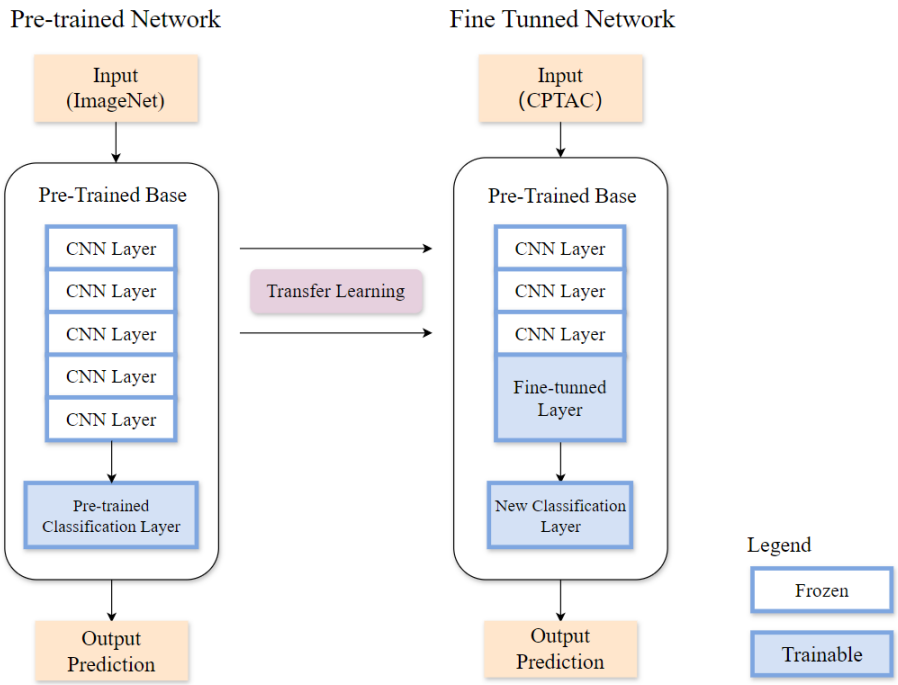


Figure 3.7: Overview of Transfer Learning with Fine Tuning.

Table 3.2: Model Training Configuration and Parameters.

Parameter	Details
Epochs	50
Batch Size	32
Input Dimensions	224 × 224 × 3 (RGB image patch)
Pretrained Layers	80% frozen, final 20% fine-tuned
Top Architecture	GlobalAveragePooling2D → Dense(512) → Dropout(0.5) → Dense(1)
Activation Functions	ReLU (Dense 512), Sigmoid (Output layer)
Regularizer	L2 regularization (λ = 0.01) on Dense(512)
Optimizer	AdaBound
Learning Rate	0.00001 (1e-5)
Loss Function	Binary Cross-Entropy
Callbacks	EarlyStopping (monitor = val_accuracy, patience = 5, restore best weights)

3.5 Esemble Learning

Various CNN architectures were used to generate diverse models, each trained on the same histopathological dataset. The top-3 and top-5 highest accuracy predictions from these individual models were then aggregated using various techniques, including simple methods like averaging and majority voting, as well as advanced techniques including stacking. This approach leveraged the strengths of each model and reduces the likelihood of overfitting or model-specific biases, leading to more accurate and reliable predictions for classifying HNSCC histopathological images.

3.5.1 Averaging

Averaging combines predictions from multiple models by computing the average of their predictions. Both unweighted and weighted averaging ensemble had been applied in this project. In unweighted averaging, all models contributed equally to the final prediction, with each model's output treated the same regardless of its individual performance. For classification tasks, this involved averaging the predicted probabilities and selecting the class with the highest average probability. In weighted averaging, predictions were averaged with each model assigned a weight based on its performance metrics, such as accuracy. Models that perform better have a greater influence on the final prediction, leading to potentially improved results. Weighted averaging allowed for a more nuanced aggregation by recognizing and leveraging the strengths of more accurate models.

3.5.2 Majority Voting

Majority voting aggregates predictions by tallying the votes each class received from all models. In this project, hard voting was employed, where each model’s prediction counted as a vote for a specific class, and the class with the highest number of votes was selected as the final prediction. This approach was both straightforward and effective, particularly when models had comparable performance. By counting votes directly, hard voting harnessed the collective judgment of multiple models to determine the most likely class. It was especially beneficial when individual models were diverse, as their varied predictions complemented each other to enhance overall classification accuracy.

3.5.3 Stacking

Stacking, or stacked generalization, is an advanced ensemble technique where multiple base models are trained, and their predictions are aggregated through a meta-model. First, a diverse set of base models is trained separately on the same dataset. The predictions from each base model are then used as input features for a meta-model. The meta-model, trained on these predictions, learns to optimally combine the base models' outputs to make the final prediction. This process allowed the meta-model to exploit the strengths of each base learner and integrate their predictions effectively. Stacking often results in superior performance compared to any single model, as the meta-model learns the best way to synthesize the diverse predictions of the base models.

3.6 Model Evaluation

3.6.1 Performance Matics

To assess the effectiveness of the classification models, several performance metrics were utilized, all derived from the confusion matrix. The confusion matrix provides a detailed comparison between the predicted and actual outcomes, capturing key elements such as True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). By analyzing the confusion matrix from each model's testing output, these performance metrics offered a comprehensive evaluation of the model's ability to classify correctly. Collectively, these metrics highlighted the model's strengths and reveal areas for improvement, providing a robust assessment of its classification performance.

3.6.1.1 Accuracy

Accuracy (Acc) is a fundamental metric calculated as the ratio of correctly classified instances (both TP and TN) to the total number of instances. It provides an overall view of model performance, reflecting how well the model distinguishes between classes across the entire dataset.

$$Acc = \frac{TP+TN}{TP+TN+FP+FN} \tag{3.2}$$

3.6.1.2 Specificity

Specificity (SP) measures the proportion of actual negatives that are correctly identified, calculated as Equation 3.3. It represents the model's ability to correctly identify negative cases. For example, in a medical test for a disease, a high specificity means that the test is good at identifying healthy individuals as negative (those without the disease), thus minimizing the chance of falsely diagnosing someone as having the disease when they actually don't. A model with high specificity is particularly valuable when the consequences of false positives are serious, as it ensures that negative cases are accurately recognized.

$$Sp = \frac{TN}{TN+FP} \tag{3.3}$$

3.6.1.3 Precision

Precision (Pr) reflects the proportion of true positive predictions among all positive predictions which calculated as Equation 3.4. Precision is crucial when the cost of false positives is high. In situations where false positives have significant consequences, such as in medical diagnoses or fraud detection, having a high precision ensures that when the model predicts a positive result, it is likely to be correct.

$$Pr = \frac{TP}{TP+FP} \tag{3.4}$$

3.6.1.4 Recall

Recall (Re), or sensitivity, shows the model’s ability to detect all positive cases, calculated as Equation 3.5. It quantifies how well the model captures all the true positives, considering both the correctly predicted positive cases (TP) and the cases that were missed (FN). For instance, in a cancer detection model, a high recall ensures that most of the actual cancer cases are identified, even if it means incorrectly classifying some healthy individuals (leading to more false positives). A low recall would mean that many positive cases (e.g., people with cancer) are not being detected, which could have serious consequences.

$$Re = \frac{TP}{TP+FN} \quad (3.5)$$

3.6.1.5 F1 Score

The F1 Score combines both precision and recall to provide a balanced measure of a model's performance, especially when there is an uneven class distribution or when both false positives and false negatives are of concern. It is calculated as Equation 3.6. A high F1 score indicates that the model performs well in terms of both precision and recall, making it a reliable measure of the model's overall classification ability.

$$F1 = 2 \times \frac{Pr \times Re}{Pr+Re} \quad (3.6)$$

3.6.1.6 Area under Curve

The Area Under the Curve (AUC) evaluates the model's ability to distinguish between classes across different thresholds, providing a summary of the model's performance. The ROC curve itself is created by plotting the True Positive Rate (Recall) against the False Positive Rate at different threshold values. By adjusting the threshold for classifying positive and negative cases, the ROC curve illustrates how well the model balances between correctly identifying true positives and minimizing false positives. The AUC represents the total area under this curve, with values ranging from 0 to 1. An AUC of 1.0 indicates perfect classification, where the model fully distinguishes between classes, while an AUC of 0.5 implies the model performs no better than random guessing. The closer the AUC value is to 1, the better the model's ability to separate the positive and negative classes.

3.6.1.7 Matthews Correlation Coefficient

Matthews Correlation Coefficient (MCC) is a comprehensive measure of the model's performance, considering all four categories of the confusion matrix, and is calculated as Equation 3.7. Unlike accuracy, which can be misleading in imbalanced datasets, MCC provides a balanced view of the model's performance by taking into account both the correct classifications and the types of errors made. MCC ranges from -1 to 1, where 1 indicates perfect

classification, 0 suggests no correlation between predictions and actual outcomes, and -1 reflects complete misclassification.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (3.7)$$

3.6.1.8 Diagnostic Odds Ratio

The Diagnostic Odds Ratio (DOR) evaluates the odds of a positive test result being correctly identified in positive cases versus negative cases, calculated as Equation 3.8. This metric provides an indication of the overall effectiveness of the diagnostic test. A higher DOR reflects improved test performance, with values above 1 indicating that the test is more effective at differentiating between positive and negative cases.

$$DOR = \frac{TP \times TN}{FP \times FN} \quad (3.8)$$

3.6.2 Tumor Prediction Heatmaps using Grad-CAM

To enhance the interpretability of the CNN model's predictions, Gradient-weighted Class Activation Mapping (Grad-CAM) was employed. This technique generates visual heatmaps that feature the regions of an image that most influence the model's decision for predicting tumors. Typically, these heatmaps use a color gradient (e.g., from blue to red) to represent intensity. Areas with higher intensity are usually shown in warmer colors like red or yellow, while areas with lower intensity are depicted in cooler colors like blue or green. Grad-CAM was applied to the final convolutional layers of each CNN model (VGG, ResNet, DenseNet, MobileNet, and EfficientNet) to visualize their focus areas.

In this project, the Grad-CAM process involved extracting the feature maps and gradients from the last convolutional layer, computing a weighted combination based on the gradients, and generating a normalized heatmap that was overlaid onto the original histopathological image for visual interpretation. An ensemble Grad-CAM approach was also implemented. For each image, Grad-CAM heatmaps were individually generated from the selected CNN models and resized to a uniform target shape. These heatmaps were then

averaged to produce a composite ensemble heatmap, which emphasizes consistent activation regions across all models. This ensemble method provides a more reliable interpretation by highlighting regions that multiple models agree are significant for classification.

3.7 Baseline Comparison Using Vision Transformer (ViT)

To establish a baseline for comparison against CNN-based models, a ViT model was implemented using the Hugging Face transformers library. The model was initialized with pretrained weights from a ViT variant and fine-tuned on the selected dataset with binary labels representing ‘Cancer’ and ‘No Cancer’. The label mapping was explicitly defined using label2id and id2label dictionaries to ensure consistent interpretation during training and inference.

Fine-tuning was carried out using the Hugging Face Trainer API with the following key training configurations: a batch size of 32, learning rate of 2e-4, and a total of 30 training epochs. The training used mixed precision (fp16) to accelerate computation and reduce memory usage. Model evaluation was performed at regular intervals, and early stopping was incorporated with a patience of 6 evaluation steps to prevent overfitting and minimize training time. The best model checkpoint was automatically selected based on validation performance.

After training, the model was evaluated on the test dataset, and predictions were generated by passing each test image through the model. The predicted class for each image was determined by selecting the class with the highest logit score from the model’s output. These predictions were then directly evaluated using performance metrics and compared with the results from the CNN-based ensemble models.

3.8 Experiment Settings

All CNN and ensemble model training, evaluation, and preprocessing tasks were conducted on a Windows 10 workstation powered by dual Intel XEON E5-2630v3 CPUs, an NVIDIA Quadro P6000 GPU, 120 GB of RAM, and CUDA version 11.4. The environment was configured with Python 3.8.19 and TensorFlow 2.5. In contrast, the ViT models were trained on Google Colab

using Python 3.11.12, PyTorch 2.6.0+cu124, and the Transformers library version 4.37.2.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Introduction

In this project, fine-tuned trainings were done with CNNs of five major family, including DenseNet, EfficientNet, MobileNet, ResNet and VGG, to demonstrate their effectiveness in histopathological image classification and contribute to the objective of enhancing diagnostic performance through deep learning. Prior to training, the histopathological whole slide images were successfully tiled into patches using a 50% tissue threshold and underwent stain normalization using the Vahadane method. The dataset was split into 80% training, 20% validation, and 10% testing, with the benign slide ratio controlled between 22% and 26% within each category. To further address data imbalance, L2 regularization was applied to the model training process for smoothing the training process and avoiding sharp weight updates that could cause overfitting.

To assess model robustness across varying visual contexts, three magnification levels (1.25×, 2.5×, and 5×) were tested. Different magnification levels provide varying tissue detail, where lower magnifications offer broader structural context, while higher magnifications reveal cellular features. Testing multiple levels helps identify which resolution yields the best diagnostic performance and supports multi-scale representation learning. A ViT model was also fine-tuned for comparison, and an ensemble voting strategy was later applied to explore performance gains through model combination. Together, this preprocessing and training pipeline demonstrated reliable performance across models and magnification levels.

4.2 Classification Result

The classification results were evaluated by comparing the performance of the proposed CNN architectures with 1.25x (Table 4.1), 2.5x (Table 4.2), and 5x (Table 4.3) magnification input using various metrics, including Acc, Sp, Pr, Re, F1, AUC, MCC, and DOR. Figure 4.1 was plotted to compare the classification performance of CNN-based models and ViT across three magnification levels

(1.25×, 2.5×, and 5×), facilitating an easy comparison of their respective accuracies.

Table 4.1: Performance Metrics Comparison of Deep Learning Architectures for 1.25x Magnification.

Architecture	Performance Metrics							
	Acc	**Sp**	**Pr**	**Re**	**F1**	**AUC**	**MCC**	**DOR**
DenseNet121	0.8903	0.7315	0.9061	0.9484	0.9268	0.8400	0.7115	50
DenseNet169	0.9065	0.7315	0.9080	0.9705	0.9382	0.8510	0.7536	90
DenseNet201	0.8975	0.8523	0.9442	0.9140	0.9288	0.8832	0.7470	61
EfficientNetB0	0.9083	0.7181	0.9045	0.9779	0.9398	0.8480	0.7587	113
EfficientNetB1	0.9281	0.7987	0.9297	0.9754	0.9520	0.8870	0.8122	**157**
EfficientNetB2	0.8867	0.6980	0.8963	0.9558	0.9251	0.8269	0.6997	50
EfficientNetB3	0.9011	0.7181	0.9037	0.9681	0.9348	0.8431	0.7387	77
EfficientNetB4	0.8849	0.6376	0.8803	0.9754	0.9254	0.8065	0.6937	70
EfficientNetB7	0.8975	0.6980	0.8977	0.9705	0.9327	0.8343	0.7287	76
MobileNet	0.8525	0.6242	0.8719	0.9361	0.9028	0.7801	0.6050	24
MobileNetV2	0.7932	0.2752	0.7874	0.9828	0.8743	0.6290	0.4068	22
ResNet50	0.8975	0.6913	0.8959	0.9730	0.9329	0.8321	0.7287	81
ResNet101	**0.9317**	0.8255	0.9382	0.9705	**0.9541**	0.8980	0.8223	156
ResNet152	0.9155	0.7584	0.9167	0.9730	0.9440	0.8657	0.7782	113
VGG16	**0.9317**	**0.8792**	**0.9556**	0.9509	0.9532	**0.9150**	**0.8266**	141
VGG19	0.9173	0.7919	0.9267	0.9631	0.9446	0.8775	0.7840	99
Ensemble-MV3[1]	**0.9514**	**0.8523**	**0.9481**	0.9877	**0.9675**	**0.9200**	0.8744	464
Ensemble-UA3[2]	**0.9514**	0.8456	0.9460	0.9902	0.9676	0.9179	**0.8746**	552
Ensemble-WA3[3]	0.9478	0.8121	0.9355	**0.9975**	0.9655	0.9048	0.8665	**1755**
Ensemble Stack3[4]	0.9460	0.8121	0.9353	0.9951	0.9643	0.9036	0.8613	875
Ensemble-MV5	0.9424	**0.8255**	**0.9391**	0.9853	0.9616	0.9054	0.8507	316
Ensemble-UA5	**0.9460**	0.8188	0.9374	0.9926	0.9642	**0.9057**	0.8609	608
Ensemble-WA5	0.9406	0.7852	0.9269	**0.9975**	0.9609	0.8914	0.8479	**1484**
Ensemble Stack5	**0.9460**	0.8121	0.9353	0.9951	**0.9643**	0.9036	**0.8613**	875
ViT	0.8687	0.7517	0.9093	0.9115	0.9104	0.8320	0.6646	31

[1] MV3 represents the ensemble model with majority volting method for Top-3 highest accuracy CNN Model

[2] UA3 represents the ensemble model with unweighted averaging Method

[3] WA3 represents the ensemble model with weighted averaging Method

[4] Stack3 represents the ensemble model with stack method

Table 4.2: Performance Metrics Comparison of Deep Learning Architectures for 2.5x Magnification.

Architecture	Performance Metrics							
	Acc	**Sp**	**Pr**	**Re**	**F1**	**AUC**	**MCC**	**DOR**
DenseNet121	0.9361	0.8344	0.9497	0.9676	0.9586	0.9010	0.8199	150
DenseNet169	0.9465	0.9094	0.9716	0.9580	0.9648	0.9337	0.8545	229
DenseNet201	0.9438	0.9224	0.9754	0.9504	0.9627	0.9364	0.8497	228
EfficientNetB0	**0.9526**	0.9159	0.9737	0.9640	**0.9688**	**0.9400**	**0.8704**	**292**
EfficientNetB1	0.9487	**0.9237**	**0.9759**	0.9564	0.9661	**0.9400**	0.8616	265
EfficientNetB2	0.9514	0.8732	0.9614	**0.9756**	0.9684	0.9244	0.8635	275
EfficientNetB3	0.9490	0.8849	0.9646	0.9688	0.9667	0.9268	0.8579	239
EfficientNetB4	0.9227	0.8642	0.9573	0.9408	0.9490	0.9025	0.7903	101
EfficientNetB7	0.9065	0.7646	0.9289	0.9504	0.9395	0.8575	0.7349	62
MobileNet	0.8772	0.6507	0.8976	0.9472	0.9218	0.7990	0.6422	33
MobileNetV2	0.8619	0.5511	0.8735	0.9580	0.9138	0.7545	0.5865	28
ResNet50	0.9453	0.8926	0.9666	0.9616	0.9641	0.9271	0.8493	208
ResNet101	0.9388	0.8140	0.9538	0.9703	0.9620	0.9301	0.8058	143
ResNet152	0.9401	0.8292	0.9486	0.9744	0.9613	0.9018	0.8303	185
VGG16	0.9444	0.8564	0.9563	0.9716	0.9639	0.9140	0.8436	204
VGG19	0.9157	0.7723	0.9317	0.9600	0.9456	0.8662	0.7597	81
Ensemble-MV3	**0.9554**	**0.9030**	**0.9700**	0.9716	**0.9708**	**0.9373**	**0.8761**	318
Ensemble-UA3	**0.9554**	0.9004	0.9693	0.9724	**0.9708**	0.9364	0.8759	318
Ensemble-WA3	0.9551	0.8875	0.9656	**0.9760**	**0.9708**	0.9317	0.8742	**321**
Ensemble Stack3	0.9548	0.8952	0.9678	0.9732	0.9705	0.9343	0.8739	310
Ensemble-MV5	**0.9609**	**0.9198**	**0.9752**	0.9736	**0.9744**	**0.9467**	**0.8918**	**423**
Ensemble-UA5	0.9523	0.8900	0.9662	0.9716	0.9689	0.9308	0.8671	277
Ensemble-WA5	0.9560	0.8926	0.9671	**0.9756**	0.9713	0.9341	0.8770	332
Ensemble Stack5	0.9600	0.9185	0.9747	0.9728	0.9738	0.9556	0.8893	403
ViT	0.9021	0.6438	0.9148	0.9674	0.9404	0.9360	0.6762	54

Table 4.3: Performance Metrics Comparison of Deep Learning Architectures for 5x Magnification.

Architecture	Performance Metrics							
	Acc	**Sp**	**Pr**	**Re**	**F1**	**AUC**	**MCC**	**DOR**
DenseNet121	0.9126	0.7584	0.9310	0.9579	0.9443	0.8581	0.7437	71
DenseNet169	0.9055	0.7389	0.9256	0.9545	0.9398	0.8467	0.7223	59
DenseNet201	0.8928	0.6375	0.9008	0.9678	0.9331	0.8027	0.6758	53
EfficientNetB0	**0.9272**	0.8295	0.9502	0.9559	**0.9530**	**0.8927**	**0.7912**	**105**
EfficientNetB1	0.9169	0.8166	0.9461	0.9464	0.9463	0.8815	0.7634	79
EfficientNetB2	0.9223	0.8235	0.9483	0.9513	0.9498	0.8874	0.7778	91
EfficientNetB3	0.9236	**0.8529**	**0.9562**	0.9444	0.9503	0.8987	0.7860	99
EfficientNetB4	0.9228	0.8189	0.9471	0.9533	0.9502	0.8861	0.7784	92
EfficientNetB7	0.9175	0.7901	0.9393	0.9549	0.9470	0.8725	0.7607	80
MobileNet	0.8759	0.5578	0.8818	0.9694	0.9235	0.7636	0.6179	40
MobileNetV2	0.8628	0.4656	0.8618	**0.9795**	0.9169	0.7225	0.5706	42
ResNet50	0.9121	0.7344	0.9251	0.9644	0.9443	0.9494	0.7399	75
ResNet101	0.9177	0.8486	0.9547	0.9380	0.9463	0.8933	0.7709	85
ResNet152	0.9119	0.7629	0.9320	0.9557	0.9437	0.8593	0.7424	69
VGG16	0.9208	0.8038	0.9430	0.9552	0.9491	0.8795	0.7711	87
VGG19	0.9162	0.8095	0.9442	0.9475	0.9459	0.8785	0.7602	77
Ensemble-MV3	0.9371	**0.8600**	**0.9589**	0.9598	0.9593	**0.9099**	0.8207	147
Ensemble-UA3	**0.9384**	0.8566	0.9580	0.9625	**0.9603**	0.9095	**0.8237**	**153**
Ensemble-WA3	0.9347	0.8161	0.9472	**0.9695**	0.9582	0.8928	0.8096	141
Ensemble Stack3	0.9360	0.8255	0.9497	0.9685	0.9590	0.8970	0.8142	145
Ensemble-MV5	**0.9394**	**0.8589**	**0.9587**	0.9631	**0.9609**	**0.9110**	**0.8265**	**159**
Ensemble-UA5	0.9367	0.8486	0.9558	0.9626	0.9592	0.9056	0.8182	144
Ensemble-WA5	0.9345	0.8141	0.9467	**0.9699**	0.9582	0.8920	0.8091	141
Ensemble Stack5	0.9365	0.8243	0.9494	0.9695	0.9594	0.8969	0.8155	149
ViT	0.8935	0.7207	0.9200	0.9443	0.9320	0.9370	0.6880	44

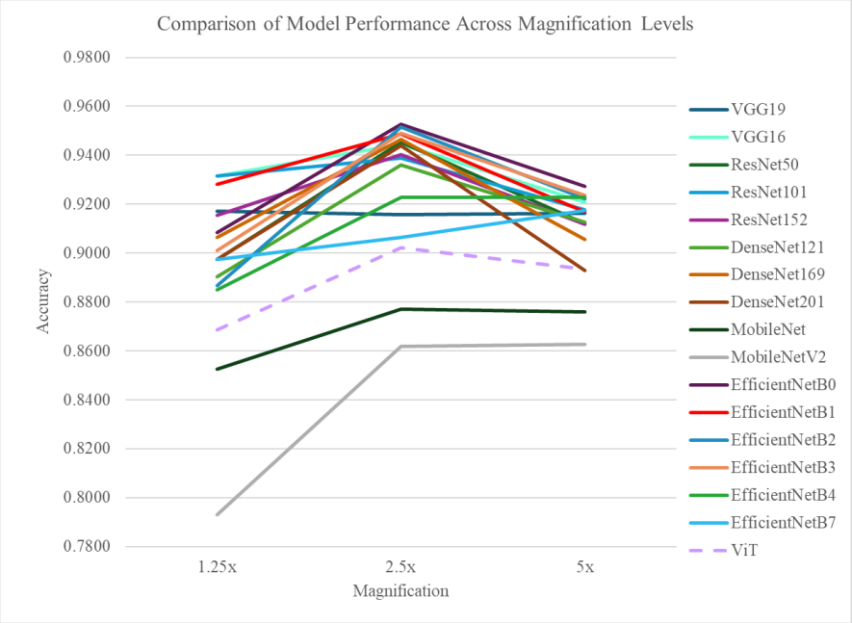


Figure 4.1: Comparison of classification accuracy for CNN-based models and ViT across three magnification levels (1.25×, 2.5×, and 5×).

4.3 Classification Performance Analysis of CNN Architectures

For **VGG** family, they demonstrated consistent and promising performance across all three magnification levels, with VGG16 standing out in particular. As shown in Table 4.4, which recorded the Top-5 performing models, VGG16 achieved the highest accuracy at 1.25× magnification with 93.17%, ranked among the Top-5 at 5× magnification with 92.08% accuracy. VGG19 also performed well, ranking fourth at 1.25× with an accuracy of 91.73%. Across all magnifications, both VGG16 and VGG19 showed strong results, although VGG16 consistently outperformed VGG19 in terms of accuracy, precision, and AUC. both variants exhibited a drop in precision and MCC at higher magnifications, indicating an increase in false positives. This suggests that at higher magnifications, the models became more aggressive in detecting cancer, leading to an increased risk of misclassifying normal tissue as malignant.

For **ResNet** family, ResNet101 generally outperformed the other variants at lower magnifications (1.25× and 2.5×), with the highest accuracy and MCC at 1.25× and strong recall at 2.5×. However, at 5×, its F1-score and MCC dropped, allowing ResNet50 to rival or even exceed its performance in some metrics. This could be related to ResNet101's depth, which allows for subtle feature extraction at moderate resolutions but becomes less efficient on high-

resolution inputs, whereas shallower models such as ResNet50 may generalize better and resist overfitting. Besides, despite being deeper, ResNet152 showed less consistency, with signs of overfitting at 2.5× (lower MCC) and a slight drop in recall at 5×. This suggests that excessive depth may introduce noise sensitivity and reduced generalization on fine-grained patches. Moreover, ResNet50, the shallower variant, performed moderately and consistently b but underperformed at 1.25× and 5×, likely due to limited capacity to capture complex features across scales. These trends suggest that ResNet101 offers an optimal trade-off between representational power and generalization in histopathological image classification, while excessively deep architectures like ResNet152 may become vulnerable to noise or reduced contextual diversity.

For **DenseNet** family, they demonstrated generally consistent performance across the three magnification levels among other CNNs, serving as average-performing models among CNNs. Across all magnification levels, DenseNet169 emerged as the most balanced variant, achieving the highest accuracy (94.65%) along with strong recall, F1-score, and MCC. In contrast, DenseNet121, the lightest variant, showed comparatively lower performance at 1.25× and 2.5×, likely due to its limited depth restricting its ability to capture rich feature hierarchies. Nevertheless, it remained competitive, with accuracy consistently above 89%, showing its efficiency despite lower complexity. DenseNet201 performed well at 2.5x due to its dense connectivity, allowing for gradient flow and feature reuse, but suffered at higher magnification. These patterns indicate that DenseNet169 offers an effective trade-off between model complexity and generalization across varying resolutions.

The **MobileNet** family delivered lightweight and efficient performance across magnifications, though generally underperformed compared to deeper networks like ResNet and DenseNet. The best result within the family was achieved by MobileNet at 2.5× magnification, with 87.72% accuracy, strong specificity, presision and MCC. In contrast, the original MobileNetv2 struggled at all scale, which may due to its architecture trades accuracy for speed, which becomes a limiting factor in tasks requiring deep visual understanding. At 1.25× magnification, both variants experienced performance drops, particularly in recall and MCC, suggesting that their shallow architectures could not effectively capture broad contextual information from low-resolution patches. Remarkably,

both models showed more stable performance at 2.5× magnification, where the moderate-frequency details appeared to align better with MobileNet's depth-wise separable convolutions. The higher-capacity model achieved 87.72% accuracy with improved precision and MCC, showing better adaptation to detailed tissue regions. However, overall variability across scales indicates that while MobileNet is computationally efficient, it lacks the representational power to consistently handle the complexity of histopathological features at varying resolutions.

Last but not least, the **EfficientNet** family showed consistently high and stable performance across all magnification levels, outperforming most other model families in overall metrics. This is evident as EfficientNet models dominated the Top-5 highest-performing models at 2.5× and also held a strong presence at 5×, outperforming most other CNN families. EfficientNetB0 standing out by achieving the highest accuracy of 95.26% at 2.5x magnification and the best overall performance across metrics. This superior result can be attributed to EfficientNet's compound scaling strategy, which uniformly balances depth, width, and input resolution, allowing deeper models to learn complex features more effectively. Notably, even smaller variants like EfficientNetB0 and B1 performed competitively at 2.5×, indicating the efficiency of the architecture regardless of model size. However, performance declined at the lower (1.25×) and higher (5×) magnifications. Deeper variants such as B3 and B4 were more affected, possibly because their increased complexity demands richer and more balanced visual information. Surprisingly, the deepest variant, EfficientNetB7, demonstrated improved performance with increasing magnification, suggesting its potential to capture finer histological details at higher resolutions. These results highlight EfficientNet's strength in handling multi-scale image resolutions while also revealing its sensitivity to suboptimal input scales, particularly for less complex tissue features or lower-resolution patterns.

In short, EfficientNet and VGG families stood out as the top-performing architectures, consistently ranking among the best models across magnification levels. EfficientNetB0 demonstrated the highest overall performance with balanced depth and efficiency, while VGG16 showed

unexpected competitiveness despite its older architecture, likely due to its simplicity and strong feature extraction at early layers.

In **comaparison among all CNNs models**, the 2.5× magnification consistently yielded the highest overall performance across most architectures variants, including ResNet, DenseNet, MobileNet, and EfficientNet. This intermediate scale appears to offer an optimal balance between local cellular detail and broader tissue context, allowing models to extract discriminative features without being overwhelmed by noise or losing critical fine-grained information. In contrast, performance at 1.25× and 5× magnifications was slightly lower. At 1.25×, the broader but coarser tissue view likely lacked the resolution needed to capture subtle morphological differences, while at 5×, although high detail is present, it may introduce noise or lead to overfitting, especially in deeper models, due to limited contextual information. These findings underscore the importance of choosing an appropriate magnification level that aligns with the model's capacity to generalize and the nature of the histopathological features.

Table 4.4: Top-5 Performance of the Pre-trained Models.

Magnification	Rank	Architectures	Accuracy
1.25x	1	VGG16	0.9317
	2	ResNet101	0.9317
	3	EfficientNetB1	0.9281
	4	VGG19	0.9173
	5	ResNet152	0.9155
2.5x	1	EfficientNetB0	0.9526
	2	EfficientNetB2	0.9514
	3	EfficientNetB3	0.9490
	4	EfficientNetB1	0.9487
	5	DenseNet169	0.9465
5x	1	EfficientNetB0	0.9272
	2	EfficientNetB3	0.9236
	3	EfficientNetB4	0.9228
	4	EfficientNetB2	0.9223
	5	VGG16	0.9208

4.4 Classification Performance Analysis of Ensemble Models

The ensemble models were ensembled according to the Top-3 and Top-5 performing models with ensemble techniques, including majority voting (MV), unweighted average (UV), weighted average (WA) and stacking.

Consider the **Top-3 CNNs models ensemble**, MV3 and UA3 provided the most consistent and balanced performance, while WA3 and Stack ensemble showed slightly lower stability, likely due to the compounding of model biases through weighting. All ensemble methods achieved their highest overall metrics at 2.5x magnification. MV3 at 2.5x magnification led with the highest accuracy (95.54%), followed closely by UA3 and WA3. This trend is mirrored in other metrics such as AUC, F1-score, precision, and MCC. It may attributed to MV3 aggregates discrete class decisions, reducing the impact of individual misclassifications, while UA3 averages softmax probabilities, smoothing prediction noise. Both are basic, but effective. These observations highlighted that simple ensemble strategies like voting and unweighted averaging can be more effective and generalizable in histopathological contexts, especially when data resolution and patch variability introduce noise or fine-grained discrepancies.

Consider the **Top-5 CNNs models ensemble**, MV5 at 2.5× magnification consistently delivered the best overall performance, achieving the highest accuracy of 96.09%, along with a precision of 97.52% and the highest MCC of 89.18%. This indicates a strong balance between sensitivity and specificity, as well as a high degree of agreement between predicted and actual classifications. While UA5 also showed competitive results, especially at 1.25× magnification, where it matched Stack5 with the highest accuracy (94.60%) and slightly outperformed in recall , making it more favorable when minimizing false negatives is crucial. But UA5’s averaging mechanism might have helped balance out extreme predictions, contributing to more stable recall, although its MCC was slightly lower.

Stack5 showed competitive and stable performance, maintaining high accuracy and MCC across all magnifications, especially at 2.5× magnification. Unlike simple voting or averaging, stacking can learn to correct individual model biases through a meta-learner, which might explain its overall reliability.

However, its precision and recall metrics were slightly less optimal compared to MV5, indicating space for improvement in capturing fine-grained decision boundaries. While WA5 generally trailed behind MV5, UA5 and Stack5, especially at lower magnifications, likely due to its reliance on weight assignments that may have disproportionately favored certain base models.

From the **observation from both Top-3 and Top-5 ensemble**, WA showed limitations, which might due to its sensitivity to weight distributions of some high accuracy but with relatively poorer sensitivity or specificity models. While UA and Stack are relatively competitive, each brings its own strengths to different situations. UA is effective at smoothing output probabilities and balancing bias, excelling in recall and providing a strong alternative, particularly when minimizing false negatives is essential. On the other hand, Stack5 captures complex inter-model dependencies, offering stable, well-rounded performance while leveraging a meta-learner's learning potential. Last but not least, MV emerged as the most balanced and consistently best performer among the ensemble models. Its simple, robust mechanism avoids extreme predictions, making it a reliable choice across all magnifications. This stability highlights its ability to maintain high accuracy and precision while mitigating fluctuations in model outputs.

4.5 Comparison of Top-Performing Models Among CNNs and Ensembles

The best-performing CNN models at 1.25x, 2.5x, and 5x magnifications were grouped as Top-CNNs and moved to this section for comparison with the ensemble models. As shown in Figure 4.2, all ensemble models outperformed the top-performing CNN model at both 1.25x and 5x magnifications. However, at 2.5x magnification, UA5 slightly underperformed compared to the CNN baseline. This might due to the ensemble averaging included predictions from weaker models at that scale, slightly diluting the performance of the top-performing CNN.

Further looking into others parameters, at **1.25x**, ensemble models consistently improved key overall metrics such as accuracy, recall, F1-score, DOR, MCC, and AUC. Remarkably, all ensemble models had improved accuracy by 1% to 2% from the top CNN model. These improvements suggest

that ensembling was effective in enhancing model generalization and sensitivity at low magnification. However, this came at the cost of reduced specificity and precision, indicating a slight increase in false positives.

At **2.5x** magnification, although the performance of UA5 was slightly lower than the CNN baseline, other core performance metrics, including accuracy, recall, F1-score, DOR, and MCC, were improved. These metrics reflect better overall classification balance and robustness, particularly in detecting positive (tumor) cases, which is crucial in cancer diagnosis.The improvements ranged an increment from 0.3% to 0.8%, suggesting that the ensemble of the top five CNN models effectively captured complementary features and reduced model variance. However, the slight drop observed in specificity and precision suggests that the model may have produced more false positives, but this trade-off led to higher recall, indicating fewer tumor cases were missed. This behavior is often preferred in clinical settings, where sensitivity is prioritized to avoid missing malignant cases.

At **5x** magnification, ensemble models performed admirably, enhancing accuracy, specificity, precision, F1-score, DOR, MCC, and AUC. Ensemble models also provide an improvement ranging from 0.7% to 1.2%, indicating a more consistent and balanced categorization at higher magnification, probably due to the ensemble's ability to filter noise and refine predictions. However, recall was reduced significantly, indicating a more conservative classification method that resulted in fewer false positives while missing some real tumor cases. The result suggests that ensembles at this resolution prioritized reliability and precision, which improves clinical applicability in situations when reducing false alarms is crucial.

Overall, while ensemble techniques enhanced several performance aspects, their effectiveness varied across magnification levels. These results should be interpreted in the context of clinical priorities, such as emphasizing higher recall to minimize missed diagnoses or higher precision to reduce false positives.

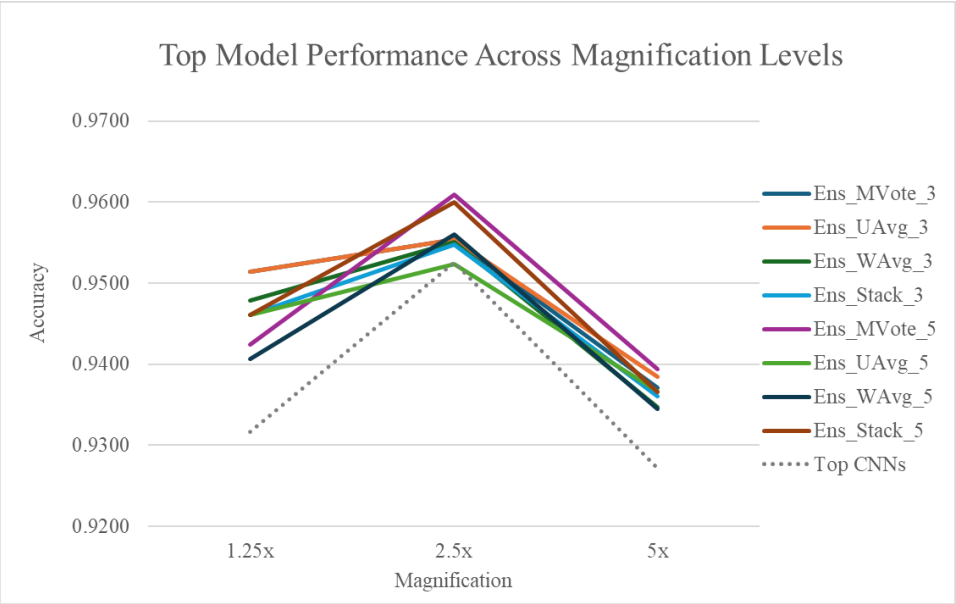


Figure 4.2: Comparison of Top-Performing Model Performance Across Magnification Levels.

4.6 Visual Interpretability Analysis of CNNs

Grad-CAM was applied to both normal and tumor tissues using the top three pretrained CNN models, including VGG, ResNet, and EfficientNet, to visualize and interpret the regions localized by each model, as shown in Figure 4.3. This approach provided a detailed visual analysis of the areas in the images that the model focused on to make its predictions. The variations in heatmap intensity values reflect the level of attention the model gives to specific areas when making decisions, with warmer colors (e.g., red and yellow) representing areas of high attention and cooler colors (e.g., blue) indicating regions of low attention.

Focusing on **benign tissue**, VGG and EfficientNet exhibit more localized activations, focusing on specific regions of the image. In contrast, ResNet show higher activation across the entire image, indicating a broader area of attention. This indicated that VGG and EfficientNet concentrate their attention on specific regions of the image for the prediction, while ResNet tends to integrate information from a broader area. These variations in activation patterns reveal how each model processes and interprets images. For instance, ResNet potentially leveraged a more comprehensive view of the image for its predictions, while EfficientNetB0 and VGG16 focused on specific regions, suggesting a more localized approach in their analysis. This difference in

activation patterns provides valuable insights into the distinct strategies each model employs for image classification.

Focusing on **tumor tissue**, all models showed significant similarity in the regions of activation, although there are some differences in color intensity, except VGG at 1.25x magnification. It obviously showed that VGG at 1.25× concentrated its attention on the corners of the image, while both ResNet and EfficientNet localized their focus toward the central region of the patch. Although all models produced correct predictions, this variation in activation patterns highlights the distinct interpretative strategies used by each model, suggesting that even with similar outcomes, their internal feature recognition processes might differ. These understandings can help guide model selection and refinement in histopathology applications, ultimately improving classification accuracy and interpretability by explaining how each model makes its decision.

Furthermore, Grad-CAM visualizations for benign tissue typically appeared predominantly blue, indicating low model activation across the image. This reflects the absence of discriminative pathological features, suggesting that the model confidently recognized the tissue as normal without focusing on any specific abnormal region. In some normal tissue patches, such as the benign tissue classified by ResNet at 2.5x magnification, Grad-CAM visualizations showed localized regions of high activation. Despite this focused attention, the overall prediction probability remained below the tumor classification threshold. This suggests that while the model identified potentially ambiguous features, they were not sufficient to override its classification of the tissue as benign. It may also reflect areas of normal histological variability or borderline features that the model considered but ultimately dismissed.

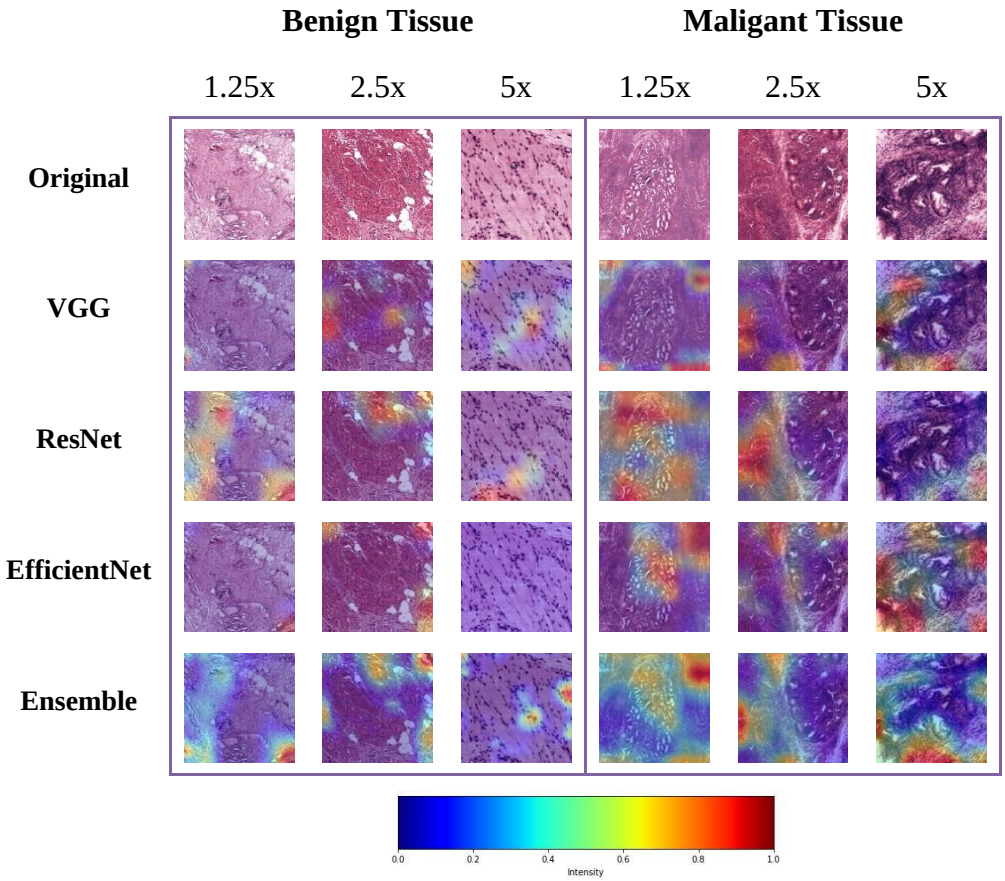


Figure 4.3: Visual Interpretability Analysis using Grad-Cam.

4.7 Performance Comparison with Vision Transformer (ViT)

To further validate the effectiveness of the proposed model, a ViT architecture was also employed as a baseline model for comparison using the same dataset, enabling performance benchmarking against non-CNN-based approaches. When compared to the fine-tuned CNN models shown in Figure 4.1, only the MobileNet family performed worse than the fine-tuned ViT. According to Figure 4.4, the best-performing CNN models, along with the ensemble models comprising three and five CNNs, all outperformed the fine-tuned ViT. This demonstrated the superior effectiveness of the proposed CNN-based approaches and the ensemble learning technique on the same dataset.

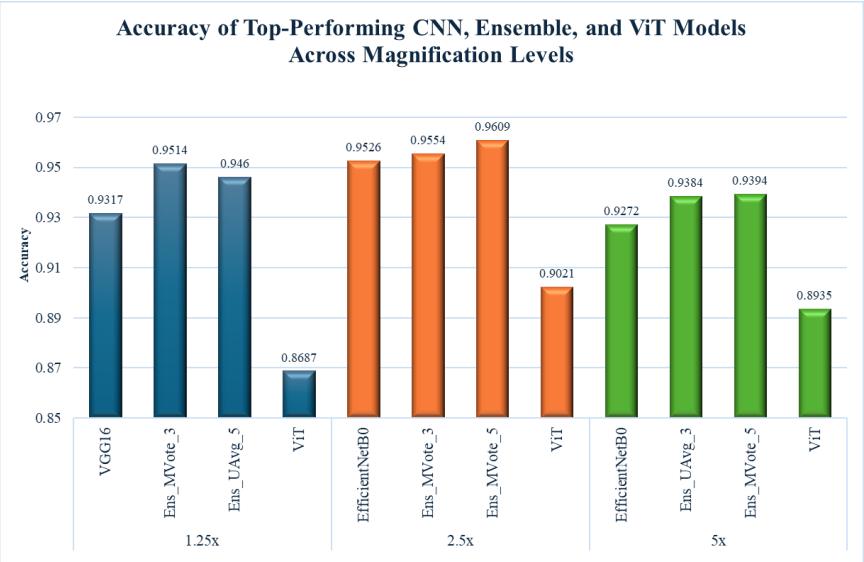


Figure 4.4: Performance Comparison among Different Models.

4.8 Challenges of the Study

One of the main challenges encountered in this study was the issue of overfitting. The CPTAC-HNSCC dataset used is relatively small for deep learning applications, which increases the risk of the models memorizing training data rather than learning generalizable patterns. Moreover, the dataset contains 122 normal slides and 268 tumor slides, resulting in a class imbalance. The models tended to favor the majority class (tumor tissue), potentially leading to overfitting, where they became overly specialized in recognizing dominant patterns while struggling to generalize to the underrepresented normal tissue in unseen data. To address this, L2 regularization was incorporated during training to reduce overfitting. While this approach helped control validation loss, a noticeable gap between validation and testing accuracy still remained, suggesting that the models had limited generalization capability and were still influenced by patterns specific to the training data.

4.9 Summary

This study proposed a robust pipeline of model training for HNSCC cancer detection. the classification performance of various CNN architectures and ensemble strategies on histopathological image patches at three different magnification levels (1.25×, 2.5×, and 5×), targeting accurate differentiation between tumor and normal tissues, which an essential task in computer-aided

cancer diagnostics. The individual CNN models displayed varied performance across magnifications, with EfficientNetB0 achieving the highest overall accuracy. EfficientNet and VGG family consistently delivered robust and balanced results, indicating their suitability as general-purpose backbones for histopathological image analysis pipelines. The ensemble models, which combined the predictions of individual CNN models, outperformed most of the individual models, highlighting the benefit of leveraging multiple architectures for improved accuracy and robustness. The MV5 ensemble model achieved the highest performance across all sectors, with an accuracy of 96.09%, outperforming both individual CNN models and other ensemble approaches.

In terms of visual interpretability, Grad-CAM visualizations indicated that the CNN models focused on regions consistent with tissue structures, providing meaningful insights into their decision-making process. The comparison of top-performing CNN models (VGG16, EfficientNetB0, and DenseNet169) revealed that while the EfficientNetB0 model excelled at higher magnifications, DenseNet169 was more robust across different magnifications.

Afterward, a performance comparison with the ViT model revealed that, while ViT delivered competitive results, CNN-based ensemble models outperformed ViT in classification tasks, particularly at higher magnifications, highlighting CNNs' superiority in histopathological image processing. Therefore, the further evident that the proposed pipeline, including image pre-processing, transfer learning with CNNs and ensembling the top performance CNNs, was effective for enhancing classification accuracy in HNSCC detection and demonstrates strong potential for broader application in digital pathology workflows. The improved performance of ensemble models over individual CNNs can be attributed to their ability to integrate diverse feature representations and decision boundaries from multiple architectures, thereby compensating for the limitations of any single model. This fusion enhances the model's robustness, reduces variance, and provides more reliable predictions, especially in complex and heterogeneous tissue samples.

Such a pipeline could be integrated into CAD systems to assist pathologists by pre-screening slides, identifying suspicious regions, or prioritizing high-risk cases, ultimately improving diagnostic efficiency and accuracy. Furthermore, the interpretability component enabled by Grad-CAM

visualizations enhances clinical trust and transparency by allowing experts to verify that model attention aligns with histologically relevant structures. The combination of ensemble learning with careful magnification handling and model interpretability tools has the potential to significantly advance AI-driven histopathological analysis, making it more reliable and suitable for real-world clinical deployment.

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusion

This study proposed a robust and systematic deep learning pipeline for the classification of head and neck squamous cell carcinoma in histopathological images, utilizing both convolutional neural network architectures and ensemble learning strategies. The dataset, sourced from the CPTAC-HNSCC cohort, was stain-normalized and processed into image patches at three magnification levels (1.25×, 2.5×, and 5×), allowing a multi-scale examination of tissue characteristics.

Individual CNN models demonstrated varied performance, with EfficientNetB0 achieving the highest accuracy among single models. Ensemble methods, particularly the MV5 ensemble, consistently outperformed individual CNN models, achieving the highest accuracy of 96.09% along with superior performance across other key metrics, including specificity (92%), precision (97%), F1-score (97%), and AUC (95%). This demonstrated the effectiveness of combining diverse CNN models to improve robustness and generalizability across magnifications. Visual interpretability through Grad-CAM provided meaningful insights into model decision-making, aligning attention maps with histological features and enhancing clinical trust. When compared with a Vision Transformer model, CNN-based ensembles delivered superior classification results. The findings validate the proposed pipeline as a reliable solution for aiding computer-aided diagnostics in histopathology.

5.2 Recommendations for future work

To improve the clinical usability and effectiveness of AI-driven histopathological investigation, future research should concentrate on creating clinically deployable AI tools that connect smoothly with digital pathology systems. This includes optimizing user interfaces for pathologists, ensuring regulatory compliance, and validating performance in prospective clinical trials. Multi-modal learning also holds significant potential for precision oncology. By

combining histopathological image features with clinical data, such as age, tumor stage, and genomic profiles, it may further enhance diagnostic accuracy and support personalized treatment planning. Incorporating larger and more heterogeneous datasets (e.g., the recent multimodal HNC dataset by Dörrich et al. (2024)) would enhance model generalization and help reduce overfitting. Implementing strategies such as class-balanced sampling, synthetic patch generation (e.g., GANs), or focal loss can help models learn better representations of minority classes. Through these advancements, future studies can bridge the gap between research and clinical translation, promoting more accurate, interpretable, and scalable cancer diagnostic tools.

REFERENCES

Abadi, J.H. and Reza, M. (2024) *Classification of Breast Cancer Cytological Images using Vision Transformers*. Concordia University. Available at: https://spectrum.library.concordia.ca/id/eprint/993825/1/JebeliHajiAbadi_MCompSc_S2024.pdf (Accessed: 12 May 2025).

Abhishek *et al.* (2024) 'Classification of Colorectal Cancer using ResNet and EfficientNet Models', *The Open Biomedical Engineering Journal*, 18(1). Available at: https://doi.org/10.2174/0118741207280703240111075752.

Ahmad, P. *et al.* (2021) 'Risk factors associated with the mortality rate of oral squamous cell carcinoma patients', *Medicine*, 100(36), p. e27127. Available at: https://doi.org/10.1097/MD.0000000000027127.

Albalawi, E. *et al.* (2024) 'Oral squamous cell carcinoma detection using EfficientNet on histopathological images', *Frontiers in Medicine*, 10. Available at: https://doi.org/10.3389/fmed.2023.1349336.

Alowais, S.A. *et al.* (2023) 'Revolutionizing healthcare: the role of artificial intelligence in clinical practice', *BMC Medical Education*, 23(1), p. 689. Available at: https://doi.org/10.1186/s12909-023-04698-z.

Ashwini, A. *et al.* (2024) 'Deep Learning ResNet-50 Framework for Automatic Early Breast Cancer Diagnosis', in *2024 International Conference on Communication, Computing and Internet of Things (IC3IoT)*. IEEE, pp. 1–5. Available at: https://doi.org/10.1109/IC3IoT60841.2024.10550264.

Barsouk, Adam *et al.* (2023) 'Epidemiology, Risk Factors, and Prevention of Head and Neck Squamous Cell Carcinoma', *Medical Sciences*, 11(2), p. 42. Available at: https://doi.org/10.3390/medsci11020042.

Boyle, P. (2024) *Is it cancer? Artificial intelligence helps doctors get a clearer picture, AAMCNews*. Available at: https://www.aamc.org/news/it-cancer-artificial-intelligence-helps-doctors-get-clearer-picture#:~:text=Trained%20on%20data%20from%20thousands,the%20human%20eye%20cannot%20detect. (Accessed: 16 September 2024).

Byfield, P., Godard, T. and Gamper, J. (2021) *StainTools*, *GitHub*. Available at: https://github.com/Peter554/StainTools (Accessed: 10 May 2025).

Chaure, N. (2024) *Variants of ResNet: A Comparative Analysis*, *Medium*. Available at: https://medium.com/@nayanchaure601/variants-of-resnet-a-comparative-analysis-63fdc1573b34 (Accessed: 18 September 2024).

Datta Gupta, K. *et al.* (2023) 'A Novel Lightweight Deep Learning-Based Histopathological Image Classification Model for IoMT', *Neural Processing Letters*, 55(1), pp. 205–228. Available at: https://doi.org/10.1007/s11063-021-10555-1.

Deebani, W. *et al.* (2025) 'Synergistic transfer learning and adversarial networks for breast cancer diagnosis: benign vs. invasive classification', *Scientific Reports*, 15(1), p. 7461. Available at: https://doi.org/10.1038/s41598-025-90288-6.

Doi, K. (2007) 'Computer-aided diagnosis in medical imaging: Historical review, current status and future potential', *Computerized Medical Imaging and Graphics*, 31(4–5), pp. 198–211. Available at: https://doi.org/10.1016/j.compmedimag.2007.02.002.

Dörrich, M. *et al.* (2024) 'A multimodal dataset for precision oncology in head and neck cancer'. Available at: https://doi.org/10.1101/2024.05.29.24308141.

Dosovitskiy, A. *et al.* (2020) 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale'.

Filippini, D.M. *et al.* (2024) 'Rare Head and Neck Cancers and Pathological Diagnosis Challenges: A Comprehensive Literature Review', *Diagnostics*, 14(21), p. 2365. Available at: https://doi.org/10.3390/diagnostics14212365.

Gopalakrishnan, V. (2023) *Image segmentation using otsu threshold selection method, Medium*. Available at: https://medium.com/@vignesh.g1609/image-segmentation-using-otsu-threshold-selection-method-856ccdacf22 (Accessed: 17 September 2024).

Guo, Z. *et al.* (2022) 'A review of the current state of the computer-aided diagnosis (CAD) systems for breast cancer diagnosis', *Open Life Sciences*, 17(1), pp. 1600–1611. Available at: https://doi.org/10.1515/biol-2022-0517.

Halalli, B. and Makandar, A. (2018) 'Computer Aided Diagnosis - Medical Image Analysis Techniques', in *Breast Imaging*. InTech. Available at: https://doi.org/10.5772/intechopen.69792.

Hava Muntean, C. and Chowkkar, M. (2022) 'Breast Cancer Detection from Histopathological Images using Deep Learning and Transfer Learning', in *2022 7th International Conference on Machine Learning Technologies (ICMLT)*. New York, NY, USA: ACM, pp. 164–169. Available at: https://doi.org/10.1145/3529399.3529426.

Hayat, M. *et al.* (2024) 'Hybrid Deep Learning EfficientNetV2 and Vision Transformer (EffNetV2-ViT) Model for Breast Cancer Histopathological Image Classification', *IEEE Access*, 12, pp. 184119–184131. Available at: https://doi.org/10.1109/ACCESS.2024.3503413.

He, K. *et al.* (2016) 'Deep residual learning for image recognition', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hoque, Md.Z. *et al.* (2024) 'Stain normalization methods for histopathology image analysis: A comprehensive review and experimental comparison', *Information Fusion*, 102, p. 101997. Available at: https://doi.org/10.1016/j.inffus.2023.101997.

Howard, A.G. (2017) 'MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications', *arXiv preprint arXiv:1704.04861* [Preprint].

Huang, G. *et al.* (2017) 'Densely connected convolutional networks', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708.

Kalra, K. (2023) *Convolutional Neural Networks for Image Classification, Medium*. Available at: https://medium.com/@khwabkalra1/convolutional-neural-networks-for-image-classification-f0754f7b94aa (Accessed: 18 September 2024).

Kanimozhi, S. and Priyadarsini, S. (2024) 'Breast Cancer Histopathological Image Classification Using CNN and VGG-19', in *2024 Third International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*. IEEE, pp. 1–6. Available at: https://doi.org/10.1109/INCOS59338.2024.10527543.

Katar, O. *et al.* (2024) 'A Novel Hybrid Model for Automatic Non-Small Cell Lung Cancer Classification Using Histopathological Images', *Diagnostics*, 14(22), p. 2497. Available at: https://doi.org/10.3390/diagnostics14222497.

Komura, D. and Ishikawa, S. (2018) 'Machine Learning Methods for Histopathological Image Analysis', *Computational and Structural Biotechnology Journal*, 16, pp. 34–42. Available at: https://doi.org/10.1016/j.csbj.2018.01.001.

Koo, J.C. *et al.* (2023) 'Non-annotated renal histopathological image analysis with deep ensemble learning', *Quantitative Imaging in Medicine and Surgery*, 13(9), pp. 5902–5920. Available at: https://doi.org/10.21037/qims-23-46.

Li, M. *et al.* (2023) 'Medical image analysis using deep learning algorithms', *Frontiers in Public Health*, 11. Available at: https://doi.org/10.3389/fpubh.2023.1273253.

Madhavan, S. and Jones, M.T. (2024) *Deep learning architectures, IBM Developer*. Available at: https://developer.ibm.com/articles/cc-machine-learning-deep-learning-architectures/ (Accessed: 10 May 2025).

Marostica, E. *et al.* (2021) 'Development of a Histopathology Informatics Pipeline for Classification and Prediction of Clinical Outcomes in Subtypes of Renal Cell Carcinoma', *Clinical Cancer Research*, 27(10), pp. 2868–2878. Available at: https://doi.org/10.1158/1078-0432.CCR-20-4119.

McCaffrey, C. *et al.* (2024) ‘Artificial intelligence in digital histopathology for predicting patient prognosis and treatment efficacy in breast cancer’, *Expert Review of Molecular Diagnostics*, 24(5), pp. 363–377. Available at: https://doi.org/10.1080/14737159.2024.2346545.

Mittal, S. (2021) ‘Ensemble of transfer learnt classifiers for recognition of cardiovascular tissues from histological images’, *Physical and Engineering Sciences in Medicine*, 44(3), pp. 655–665. Available at: https://doi.org/10.1007/s13246-021-01013-2.

Momentum, M. (2022) *A Brief History of Vision Transformers: Revisiting Two Years of Vision Research, Medium*. Available at: https://medium.com/merantix-momentum-insights/a-brief-history-of-vision-transformers-revisiting-two-years-of-vision-research-26a6bd3251f3 (Accessed: 12 May 2025).

Moscalu, M. *et al.* (2023) ‘Histopathological Images Analysis and Predictive Modeling Implemented in Digital Pathology—Current Affairs and Perspectives’, *Diagnostics*, 13(14), p. 2379. Available at: https://doi.org/10.3390/diagnostics13142379.

Muñoz-Aguirre, M. *et al.* (2020) ‘PyHIST: a histological image segmentation tool’, *PLoS computational biology*, 16(10), p. e1008349. Available at: https://doi.org/https://doi.org/10.1371/journal.pcbi.1008349.

Najjar, R. (2023) ‘Redefining Radiology: A Review of Artificial Intelligence Integration in Medical Imaging’, *Diagnostics*, 13(17), p. 2760. Available at: https://doi.org/10.3390/diagnostics13172760.

National Cancer Institute (2021) *Head and Neck Cancers*. Available at: https://www.cancer.gov/types/head-and-neck/head-neck-fact-sheet (Accessed: 9 May 2025).

National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC) (2018) *The Clinical Proteomic Tumor Analysis Consortium Head and Neck Squamous Cell Carcinoma Collection (CPTAC-HNSCC) (Version 19)*, *The Cancer Imaging Archive*. Available at: https://doi.org/10.7937/K9/TCIA.2018.UW45NH81.

Noaman, N.F. *et al.* (2024) ‘Advancing Oncology Diagnostics: AI-Enabled Early Detection of Lung Cancer Through Hybrid Histological Image Analysis’, *IEEE Access*, 12, pp. 64396–64415. Available at: https://doi.org/10.1109/ACCESS.2024.3397040.

Ong, W. *et al.* (2023) ‘Application of Machine Learning for Differentiating Bone Malignancy on Imaging: A Systematic Review’, *Cancers*, 15(6), p. 1837. Available at: https://doi.org/10.3390/cancers15061837.

Ozdemir, B. and Pacal, I. (2025) ‘A robust deep learning framework for multiclass skin cancer classification’, *Scientific Reports*, 15(1), p. 4938. Available at: https://doi.org/10.1038/s41598-025-89230-7.

Patheda, V.R. *et al.* (2025) 'A Robust Hybrid CNN+ViT Framework for Breast Cancer Classification Using Mammogram Images', *IEEE Access*, 13, pp. 77187–77195. Available at: https://doi.org/10.1109/ACCESS.2025.3563218.

Phan, N.N. *et al.* (2021) 'Prediction of Breast Cancer Recurrence Using a Deep Convolutional Neural Network Without Region-of-Interest Labeling', *Frontiers in Oncology*, 11, p. 734015. Available at: https://doi.org/https://doi.org/10.3389/fonc.2021.734015.

Potsangbam, J. and Shuleenda Devi, S. (2024) 'Classification of Breast Cancer Histopathological Images Using Transfer Learning with DenseNet121', *Procedia Computer Science*, 235, pp. 1990–1997. Available at: https://doi.org/10.1016/j.procs.2024.04.188.

Rizzo, P.C. *et al.* (2022) 'Technical and Diagnostic Issues in Whole Slide Imaging Published Validation Studies', *Frontiers in Oncology*, 12. Available at: https://doi.org/10.3389/fonc.2022.918580.

Robinson, J. (2024) *Understanding Cancer -- Diagnosis and Treatment*, *WebMD*. Available at: https://www.webmd.com/cancer/understanding-cancer-treatment (Accessed: 16 September 2024).

Sebastian, A.M. and Peter, D. (2022) 'Artificial Intelligence in Cancer Research: Trends, Challenges and Future Directions', *Life*, 12(12), p. 1991. Available at: https://doi.org/10.3390/life12121991.

Setiawan, W., Pramudita, Y.D. and Mulaab (2024) 'Transfer learning VGG for histopathological lung cancer image classification', in, p. 030009. Available at: https://doi.org/10.1063/5.0222719.

Shah, D. (2022) *Vision Transformer: What It Is & How It Works [2024 Guide]*, *V7 Newsletter*. Available at: https://www.v7labs.com/blog/vision-transformer-guide (Accessed: 12 May 2025).

Shajun Nisha, S. and Nagoor Meeral, M. (2021) 'Applications of deep learning in biomedical engineering', in *Handbook of Deep Learning in Biomedical Engineering*. Elsevier, pp. 245–270. Available at: https://doi.org/10.1016/B978-0-12-823014-5.00008-9.

Shen, H. *et al.* (2023) 'An efficient context-aware approach for whole-slide image classification', *iScience*, 26(12), p. 108175. Available at: https://doi.org/10.1016/j.isci.2023.108175.

Shoeibi, A. *et al.* (2022) 'An overview of deep learning techniques for epileptic seizures detection and prediction based on neuroimaging modalities: Methods, challenges, and future works', *Computers in Biology and Medicine*, 149, p. 106053. Available at: https://doi.org/10.1016/j.compbiomed.2022.106053.

Simonyan, K. and Zisserman, A. (2014) 'Very Deep Convolutional Networks for Large-Scale Image Recognition', *arXiv preprint arXiv:1409.1556* [Preprint].

Spectrum AI (2024) *Artificial Intelligence in Medical Imaging*. Available at: https://www.spectral-ai.com/blog/artificial-intelligence-in-medical-imaging/#:~:text=The%20system%20reduced%20false%20positives,outcomes%20based%20on%20imaging%20data. (Accessed: 16 September 2024).

Stenson, K.M. (2025) *Epidemiology and risk factors for head and neck cancer*, *UpToDate*. Available at: https://www.uptodate.com/contents/epidemiology-and-risk-factors-for-head-and-neck-cancer (Accessed: 9 May 2025).

Tan, M. (2019) 'Efficientnet: Rethinking model scaling for convolutional neural networks', *arXiv preprint arXiv:1905.11946* [Preprint].

Tiwari, A. *et al.* (2025) 'The current landscape of artificial intelligence in computational histopathology for cancer diagnosis', *Discover Oncology*, 16(1), p. 438. Available at: https://doi.org/10.1007/s12672-025-02212-z.

Vahadane, A. *et al.* (2016) 'Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images', *IEEE Transactions on Medical Imaging*, 35(8), pp. 1962–1971. Available at: https://doi.org/10.1109/TMI.2016.2529665.

Vijay, P.P. and Patil, N.C. (2016) 'Gray scale image segmentation using OTSU Thresholding optimal approach.', *Journal for Research*, 2(05).

Wang, J. *et al.* (2025) 'Artificial intelligence in cancer pathology: Applications, challenges, and future directions', *Cytojournal*, 22, p. 45. Available at: https://doi.org/10.25259/Cytojournal_272_2024.

Wright, J. (2021) *The emotional impact of a cancer misdiagnosis*, *Gadsby Wicks*. Available at: https://www.gadsbywicks.co.uk/insights/misdiagnosis/the-emotional-impact-of-a-cancer-misdiagnosis (Accessed: 16 September 2024).

Yang, G., Luo, S. and Greer, P. (2025) 'Boosting Skin Cancer Classification: A Multi-Scale Attention and Ensemble Approach with Vision Transformers', *Sensors*, 25(8), p. 2479. Available at: https://doi.org/10.3390/s25082479.

Yang, Y. (2017) 'Ensemble Learning', in *Temporal Data Mining Via Unsupervised Ensemble Learning*. Elsevier, pp. 35–56. Available at: https://doi.org/10.1016/B978-0-12-811654-8.00004-X.

Yong, M.P. *et al.* (2023) 'Histopathological Gastric Cancer Detection on GasHisSDB Dataset Using Deep Ensemble Learning', *Diagnostics*, 13(10), p. 1793. Available at: https://doi.org/10.3390/diagnostics13101793.

Zheng, Y. *et al.* (2023) 'Application of transfer learning and ensemble learning in image-level classification for breast histopathology', *Intelligent Medicine*, 3(2), pp. 115–128. Available at: https://doi.org/10.1016/j.imed.2022.05.004.

APPENDICES

Appendix A: Confusion Metrics

Table A-1: Confusion Matrix of Models at 1.25× Magnification.

Architecture	TN	FP	FN	TP
VGG16	131	18	20	387
VGG19	118	31	15	392
ResNet50	103	46	11	396
ResNet101	123	26	12	395
ResNet152	113	36	11	396
DenseNet121	109	40	21	386
DenseNet169	109	40	12	395
DenseNet201	127	22	35	372
MobileNet	93	56	26	381
MobileNetV2	41	108	7	400
EfficientNetB0	107	42	9	398
EfficientNetB1	119	30	10	397
EfficientNetB2	104	45	18	389
EfficientNetB3	107	42	13	394
EfficientNetB4	95	54	10	397
EfficientNetB7	104	45	12	395
Ens_MVote_3	127	22	5	402
Ens_UAvg_3	126	23	4	403
Ens_WAvg_3	121	28	1	406
Ens_Stack_3	121	28	2	405
Ens_MVote_5	123	26	6	401
Ens_UAvg_5	122	27	3	404
Ens_WAvg_5	117	32	1	406
Ens_Stack_5	121	28	2	405
ViT	112	37	36	371

Table A-2: Confusion Matrix of Models at 2.5× Magnification.

Architecture	TN	FP	FN	TP
VGG16	662	111	71	2429
VGG19	597	176	100	2400
ResNet50	690	83	96	2404
ResNet101	569	130	82	2682
ResNet152	641	132	64	2436
DenseNet121	645	128	81	2419
DenseNet169	703	70	105	2395
DenseNet201	713	60	124	2376
MobileNet	503	270	132	2368
MobileNetV2	426	347	105	2395
EfficientNetB0	708	65	90	2410
EfficientNetB1	714	59	109	2391
EfficientNetB2	675	98	61	2439
EfficientNetB3	684	89	78	2422
EfficientNetB4	668	105	148	2352
EfficientNetB7	591	182	124	2376
Ens_MVote_3	698	75	71	2429
Ens_UAvg_3	696	77	69	2431
Ens_WAvg_3	686	87	60	2440
Ens_Stack_3	692	81	67	2433
Ens_MVote_5	711	62	66	2434
Ens_UAvg_5	688	85	71	2429
Ens_WAvg_5	690	83	61	2439
Ens_Stack_5	710	63	68	2432
ViT	450	249	90	2674

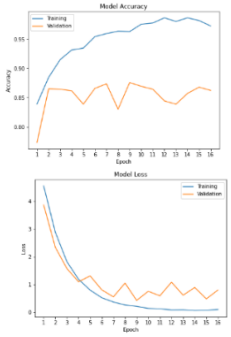
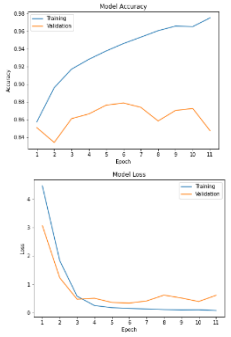
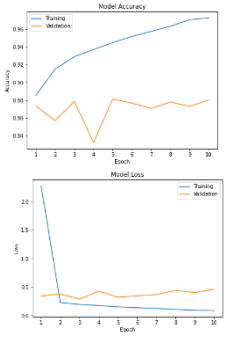
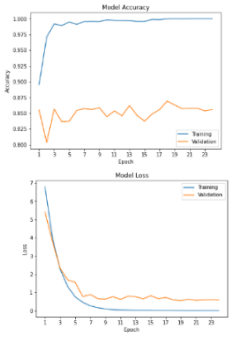
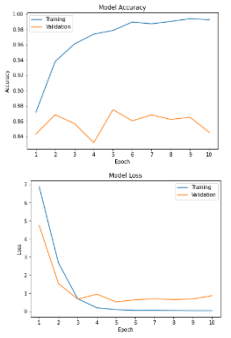
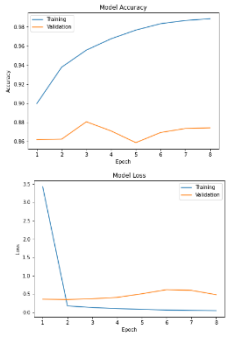
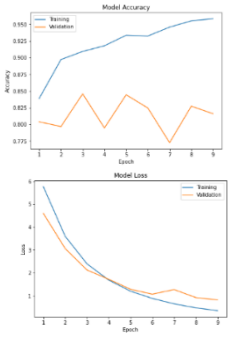
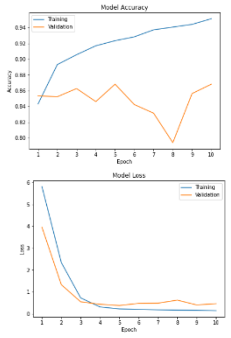
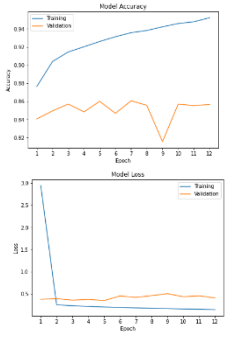
Table A-3: Confusion Matrix of Models at 5× Magnification.

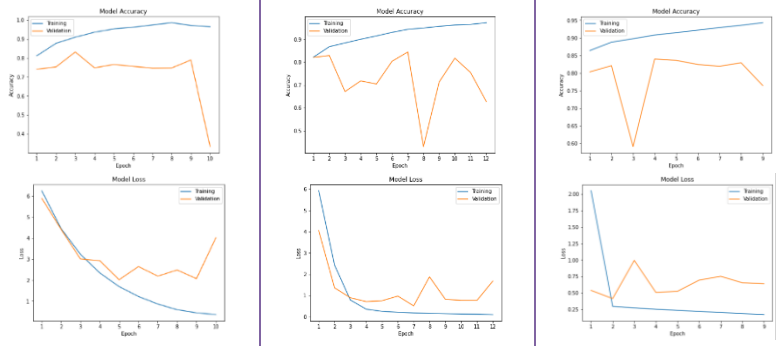
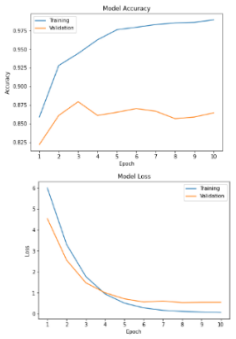
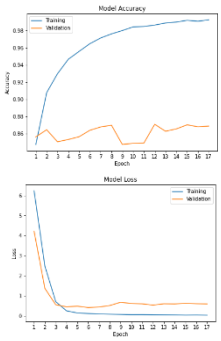
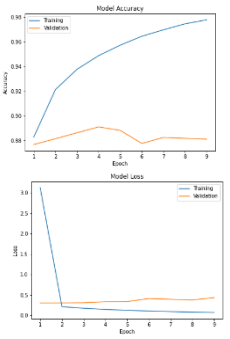
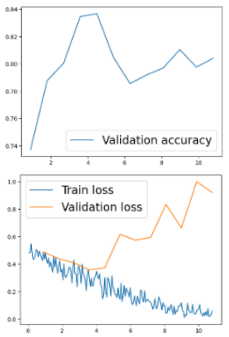
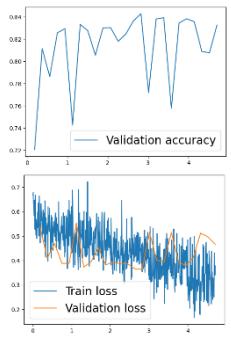
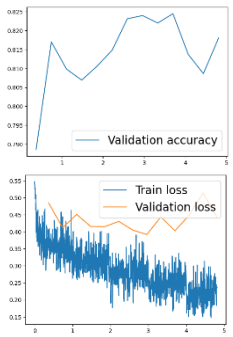
Architecture	TN	FP	FN	TP
VGG16	2814	687	534	11376
VGG19	2834	667	625	11285
ResNet50	2571	930	424	11486
ResNet101	2971	530	739	11171
ResNet152	2671	830	528	11382
DenseNet121	2655	846	501	11409
DenseNet169	2587	914	542	11368
DenseNet201	2232	1269	383	11527
MobileNet	1953	1548	365	11545
MobileNetV2	1630	1871	244	11666
EfficientNetB0	2904	597	525	11385
EfficientNetB1	2859	642	638	11272
EfficientNetB2	2883	618	580	11330
EfficientNetB3	2986	515	662	11248
EfficientNetB4	2867	634	556	11354
EfficientNetB7	2766	735	537	11373
Ens_MVote_3	3011	490	479	11431
Ens_UAvg_3	2999	502	447	11463
Ens_WAvg_3	2857	644	363	11547
Ens_Stack_3	2890	611	375	11535
Ens_MVote_5	3007	494	440	11470
Ens_UAvg_5	2971	530	446	11464
Ens_WAvg_5	2850	651	358	11552
Ens_Stack_5	2886	615	363	11547
ViT	2523	978	663	11247

Appendix B: Accuracy and Loss Curves

The accuracy and loss curves were sampled and displayed from one representative model of each architecture family.

Table B-1: Accuracy and Loss Curves Across Magnification.

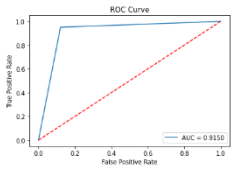
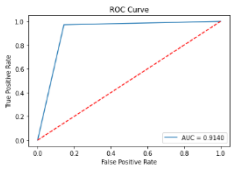
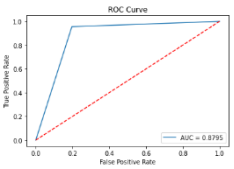
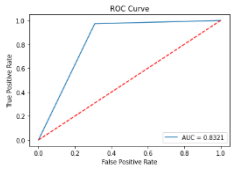
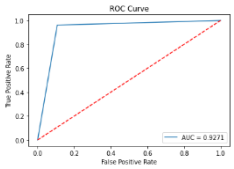
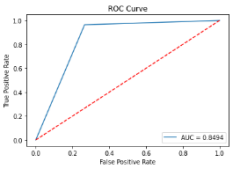
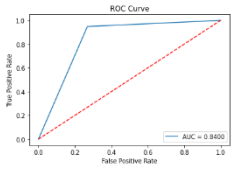
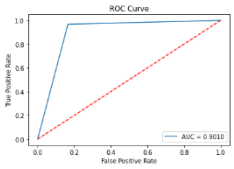
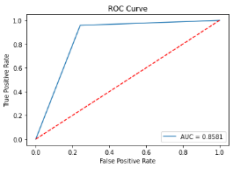
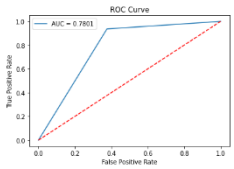
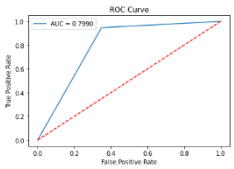
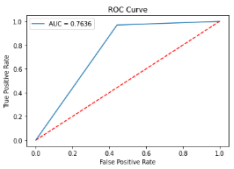
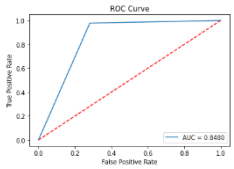
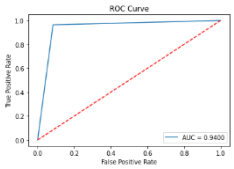
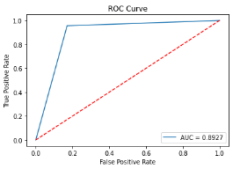
	1.25x	2.5x	5x
VGG16			
ResNet50			
DenseNet121			

	1.25x	2.5x	5x
MobileNet			
EfficientNetB0			
ViT			

Appendix C: ROC Curves

The ROC curves were sampled and displayed from one representative model of each architecture family.

Table C-1: ROC Curves Across Magnification.

	1.25x	2.5x	5x
VGG16			
ResNet50			
DenseNet121			
MobileNet			
EfficientNetB0			
ViT