Hybrid CNN–Vision Transformer for Non-Small Cell Lung Cancer Recurrence Prediction Using CT Scans

By

Sumya Akter
242-25-026

This Report Presented in Partial Fulfillment of the Requirements for
The Degree of **Masters of Science in Computer Science and Engineering**

Supervised By

Dr. S. M Aminul Haque
Professor and Associate Head
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised By

Abdus Sattar
Associate Professor Director, MSc.
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
Dhaka, Bangladesh

September, 2025

APPROVAL

This Project/Thesis titled **"Hybrid CNN-Vision Transformer for Non-Small Cell Lung Cancer Recurrence Prediction Using CT Scans"**, submitted by **Sumya Akter**, ID No: **242-25-026** to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of M.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 13-09-2025.

BOARD OF EXAMINERS

Chairman

Dr. S.M Aminul Haque
Professor & Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Ms. Nazmun Nessa Moon
Associate Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Dr. Md Alamgir Kabir
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Mr. Sadat Hossain,
Data Scientist,
Risk Management Division,
BRAC Bank Limited

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Dr. S. M. Aminul Haque**, **Professor & Associate Head**, Department of Computer Science and Engineering, Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Dr. S. M. Aminul Haque
Professor & Associate Head
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised by:

Abdus Sattar
Associate Professor & Director, MSc.
Department of Computer Science and Engineering
Daffodil International University

Submitted by:

Sumya Akter

Sumya Akter
Student ID: 242-25-026
Department of Computer Science and Engineering
Daffodil International University

ACKNOWLEDGEMENT

This work would not have been possible without the support and contributions of many individuals over the past two semesters. We are deeply grateful to all who have helped us in one way or another.

First, we express our heartfelt thanks and gratefulness to the almighty for His divine blessing making it possible for us to complete the **Hybrid CNN–Vision Transformer for Non-Small Cell Lung Cancer Recurrence Prediction Using CT Scans"** successfully.

We are grateful and wish our profound indebtedness to **Dr.S.M. Aminul Haque**, **Professor Associate Head**, Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Deep knowledge and keen interest of our supervisor in the field of Computer Vision to carry out this project. It is his un wearied forbearance and the wise counsel of the school, his constant good will, his constant and active direction, his critical remarks, which have taught so much, his good advice, his reading of so many bad drafts, his correction throughout, at every stage, which have made it possible to accomplish this projec.

I would like to express my heartfelt gratitude to **Dr. Sheak Rashed Haider Noori, Head of the Department of CSE**, for his kind assistance in completing our project, as well as to the other faculty members and staff of the CSE department at Daffodil International University.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

Artificial Intelligence (AI) in an analysis of computed tomography (CT) images. has made a major development in computer-assisted diagnosis (CAD) of lung. cancer (LC). However, challenges such as complex anatomical structures, over- dynamic positions of maladies, lapping radiographic patterns, and dynamic locations of abnormalities make. it hard to derive useful features on successful CAD implementation.Lung cancer is a major cause of disability on a global scale. Timely diagnosis is essential to. avert the development and develop special treatment plans. This study Introduces Hybrid-RViT, a hybrid neural network that combines the local. texture extraction capability of convolutional neural networks (CNNs) with the global context modeling strength of Vision Transformers (ViTs). The framework introduces three innovations: (1) Contrast Limited Adaptive Histogram Equalization (CLAHE) to increase local contrast, (2) Intensity-Based Patch Tokeniza- to divide CT scans into low-, mid-, and high-intensity images, tion (IBPT). converting a 12 channel composite input to the transformer, and (3) Smooth- based data augmentation- Gaussian blur, median filtering, bilat- eral smoothing—to increase training diversity, simulate inter-scanner variabil- ity, and enhance resistance to image noise and increasing the dataset. from 2,500 to 6,000 CT scans. Hybrid-RViT was compared to the baseline. ViT-16 and hybrid implementations with EfficientNetB0, MobileNetV2, ResNet50, and. ResNet18 backbones. Findings indicated that Hybrid-RViT using ResNet18 achieved. 97.12% accuracy, 0.97 macro-average F1-score, and 0.9982 ROC-AUC, represent- increasing an absolute and a relative accuracy of 19.28 and 24.77 percent, respectively, over the base- line Balanced ViT. All hybrid models attained perfect precision and recall of. Malignant cases, whose differentiation of Normal and has been greatly enhanced. Benign cases.The interpretable visual images given by Grad-CAM and ViT attention maps. explanations, and real-time deployment was supported by TensorFlow Lite optimization. on mobile and edge devices. The suggested framework provides a strong, cross-border. pretable, and clinically viable AI-assisted solution for lung cancer screening in low-resource settings.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Lung cancer is a leading killer type of cancer in the world and is typified with over 2.2 million cases and 1.8 million deaths annually as indicated by World health Organization [1]. Approximately 85 percent of all lung cancer Cases occur in Non-Small Cell Lung Cancer (NSCLC), and the reoccurrence rate following the surgery is fairly elevated, ranging between 30 and 55 percent [2]. Diagnosis/early detection and good prediction of recurrence is timely and of importance with regards to improving patient outcomes and shaping the personal approach to treatment.

The most important diagnostic option on pulmonary abnormality is medical imaging, especially computed tomography (CT) scans. Nonetheless, interpretation still relies on the experience of radiologists, which creates difficulties such as inter-observer variability, subjectivity, and delays in diagnosis - especially in the resource-limited healthcare systems where shortage of specialists is an issue [3].

The more recent advances in artificial intelligence (AI) specifically in the area of deep learning have demonstrated potential in automated medical image analysis. CNNs have been shown to be very successful in the local texture characteristics (tumor boundaries, mini nodules, and fine tissue patterns) acquisition.Nevertheless, CNNs usually have difficulties in capturing global contextual associations, which are critical in distinguishing the visually similar benign and malignant lesions. Conversely, Vision Transformers (ViTs) formulate long-range relationships much better than when required to extract fine-grained local information, particularly in noisy or low contrast medical data. The mutual strengthening of CNNs and ViTs leaves an opportunity to adopt a hybrid approach. The proposed Hybrid-RViT model adds their rich local features representation to CNN and their broader spatial representation to ViT in an attempt to obtain a more compound and versatile feature representation. The current architectures were not spared of the constraint of quality and size of datasets though. The datasets are small and skewed, and the low contrasts

in CT scans can prevent the appearance of features, which is why in this paper contrast-limited adaptive histogram equalization (CLAHE) to enhance the local contrast and the proposed intensity-based patch tokenization (IBPT) to define the level of intensity are proposed. Moreover, a data augmentation method Smooth is employed to boost a total of 6,000 balanced images because all the data comprise an equal Representation of normal, benign, and malignant cases. This does not only improve model generalization but also renders the proposed system useful in the real-life diagnosis setting.

1.2 Motivation

The common cause of cancer-related deaths in the world is lung cancer. According to World statistics, an estimated 2 million people are newly infected and 1.8 million die every year worldwide. The World Health Organization [1]. In general, 85 percent of the total cases of lung cancer are Non-Small Cell Lung Cancer (NSCLC). Surgery alone is not a viable long-term solution with the Journal of the Malaysian Medical Association (and other medical journals) all reporting a rate of between 20 and 70 percent post-surgical recurrence rate and causes in company with the causes of all lung cancer (about 85 percent). The numbers go up to 55 percent [2]. It is of paramount importance that the recurrence is predicted and diagnosed early. Capability to measure patient outcome and to propel personalized treatment plans. Medical imaging, Computed tomography (CT) seems to be the most common diagnostic mod-integer Relying on the physical examination, which is the main diagnostic method of diagnosing this disease, appears the most commonly used. They occurred as a sign of pulmonal abnormalities. Interpretation is, however, too much reliant. This causes problems like inter-observer variability and subject-experience because of the implication of rad-expertise. inappropriate levels of judgment and diagnostic delays - especially a problem in underresourced settings Even more than that, healthcare systems are characterized by shortages of specialty .

1.3 Research Questions

To address the limitations of existing lung cancer diagnostic systems and to explore the benefits of hybrid CNN–Transformer architectures, this study seeks to answer the following research questions:

1. How can a hybrid CNN–Vision Transformer architecture effectively combine local and global feature representations to improve lung cancer classification performance?
2. Does Intensity-Based Patch Tokenization (IBPT) enhance the Vision Transformer's ability to identify subtle lesion patterns in CT scans compared to standard patch embedding?
3. What is the impact of applying CLAHE and Smooth data augmentation on the

model's robustness, generalization, and performance in imbalanced medical imaging datasets?

4. How does the proposed Hybrid-RViT compare to other hybrid variants (e.g., CNN + ViT combinations with EfficientNetB0, MobileNetV2, and ResNet50) in terms of accuracy, F1-score, and ROC-AUC?

5. Can the proposed framework be optimized for mobile deployment via TensorFlow Lite while retaining high accuracy and interpretability using Grad-CAM and attention map visualizations?

1.4 Objectives

To overcome these barriers, this study introduces Contrast Limited Adaptive Histogram Equalization (CLAHE) for enhanced local contrast and Intensity-Based Patch Tokenization (IBPT) for intensity-aware patch embedding. Furthermore, Smooth data augmentation is applied to expand the dataset from 2,500 to 6,000 balanced images, ensuring equal representation of Normal, Benign, and Malignant cases. This not only improves model generalization but also ensures the proposed system remains effective in real-world diagnostic environments. To promote real-world applicability, the final model is also quantized and exported into a lightweight format (e.g., TensorFlow Lite) for deployment on mobile and embedded platforms. Key Contributions

1. We introduce a hybrid ResNet-ViT architecture that leverages the complementary strengths of CNNs and Transformers for robust lung disease classification.

2. We propose Intensity-Based Patch Tokenization (IBPT) to guide attention toward low-, mid-, and high-intensity regions, improving feature discrimination in clinical images.

3. We integrate CLAHE-based contrast enhancement as a core augmentation strategy to improve model generalization on low-quality medical images.

4. We incorporate Grad-CAM for visual explanation of model predictions, increasing transparency and clinical trust.

5. We demonstrate the deployment potential of the trained model via conversion to a mobile-ready, lightweight format for real-time inference.

1.5 Proposed Solution

To fill in these gaps, we present a new hybrid deep learning architecture called Hybrid-RViT by fusing the local feature learning behavior of ResNet-18 with the global attention modeling nature of a Vision Transformer (ViT). In our approach there are two main innovations: To improve the visibility of local features during preprocessing, the CLAHE is

exploited, as a contrast-enhancing approach to medical images. The proposed Intensity-Based Patch Tokenization (IBPT) maximises the discriminative learning of the ViT to tokens of different intensities. The model creates three other views of each picture through division into areas of low, medium and high strength, according to grayscale levels. These intensity-specific images concatenated with the original image provide a 12-channel compositing image which is patchified and passed to the Transformer encoder. This inspires the model to have a broader spectrum of tissue densities as well as the pixel intensity to behave like radiologists to view the image in a contrast.

1.6 Chalange

Although deep learning demonstrated positive outcomes when used in the study of lung cancer diagnosis, it still has some drawbacks that hinder the effectiveness of hybrid CNN-ViT models and its application in practice. The main issue is the fact that, the real-world CT scans are low contrast and a high amount of noise; the details of the tumor outlines/tissue patterns are in a manner lost by overlying structures and invariant quality of the scanner. This makes the extraction of reliable features in the deep models more difficult to support the classification. The other issue is the absence of emphasis on the variations of the intensity on the various tissues on the lung which in most models is given equal grounds and does not show the clinical relevance of the variations in densities that can assist in the identification of the emerging pathological changes. Moreover, the interpretability of AI-driven systems is a key barrier; in as much as high precision is preferable, the clinicians require clear models that they can utilize to justify what they made their decision based on. In addition, the interpretability of AI-based systems has remained a severe obstacle; although a high level of accuracy is desirable, clinicians need to have some model interpretability in order to have a visual explanation of why the model made a particular prediction so that they can establish trust in AI-made decisions. Moreover, this process can be complicated by the imbalance of datasets and a shortage of annotated CT scans, that is why this may cause overfitting, and the resulting model cannot be successfully applied in a variety of clinical settings. Lastly, the advanced hybrid architectures have high computational complexity that limit their deployment in real-time or mobile applications, in a healthcare scenario whereby high-performance machines are scarce in resource-constrained health systems. Addressing these limitations is critical in coming up with a powerful, explainable and lightweight AI model that can support radiologists in proper and fast lung cancer diagnosis.

1.7 Organization of the Report

The rest of this report has been divided into six chapters. Chapter 1 presents the background, motivation, problem statement, objectives, and the general contributions of the work. Chapter 2 discusses the literature on deep learning applied to the analysis of lung

CT, and it specifically addresses CNN-based models, Vision Transformer (ViTs), and hybrid and explainable AI (XAI) methods. Chapter 3 introduces the proposed methodology, such as the two-stage preprocessing pipeline, data preparation, data augmentation techniques, and the development of the hybrid CNNViT structure that makes use of CLAHE to process the ResNet branch and Intensity-Based Patch Tokenization (IBPT) to process the ViT branch. The chapter 4 explains the experimental configuration and the analysis of results, including metrics in evaluation, a comparison with other models, an analysis by a class, and an analysis of the computational efficiency. The chapter 5 pays attention to the introduction of XAI to interpret the model and outlines the implementation of the new system to run on mobile devices to provide real-time diagnostic assistance. Lastly, Chapter 6 will wrap up the report in a summary of the key findings, contribution of the research, and future research directions, including large-scale clinical validation and integration of multimodal data.

Chapter 2

Background

2.1 Overview

The history of the literature provides a knowledge foundation on the application of artificial intelligence in the lung cancer diagnosis. Previous works have explored fully CNNs to capture surrounding textures and fully ViTs to capture global dependencies and others have been introduced in the recent times to combine the two forms. These works show strengths in both architectures but also display new weaknesses such as a lack of interpretation of how they achieved their conclusions (poor explainability), no intensity-based preprocessing, and the inability to model low quality CTs. There is need to review these contributions because they are crucial in terms of what has been accomplished and where there exist gaps. In our part, we critically reviewed the literature of the past year 2022-2025 to present latest advancements that have been made. We contrasted different hybrid CNN influx-transformer (ViT) models, and their performance and problems with deployment. Due to the process, we have placed our research within the overall scope, stating how our Hybrid-RViT fills the missing gap on contrast enhancement, intensity-based patch-wise tokenization, explanation, and mobile applicability. The literature review, therefore, cannot only provide the insight on the state-of-the-art improvements but is quite persuasive in light of the novelties and needs of our proposed approach.

2.2 Literature Review

Recent advances in artificial intelligence (AI) and deep learning have significantly impacted lung cancer (LC) diagnosis, particularly through automated analysis of imaging modalities such as computed tomography (CT), magnetic resonance imaging (MRI), and histopathology. Javed et al. [2] provided a comprehensive systematic review exploring deep learning applications—including convolutional neural networks (CNNs) and Vision Transformers (ViTs)—across diverse imaging modalities. Their analysis highlights the evolution of AI-based lung cancer detection from 2015 to 2024, emphasizing improved diagnostic accuracy through automated feature extraction. A number of studies have

concentrated on the implementation of Vision Transformer (ViT) architectures to lung images. Ahmad et al.[3] used ViT and Swin models to classify chest X-rays as normal, lung opacities and pneumonia with 99 percent binary classification and 95.25 percent multicast classification. Jain et al. [4] compared standard CNNs, ResNet, and ViT models and found that hybrid ViT-ResNet networks are able to more effectively capture more complex spatial relationships further improving multi-class lung disease classification. On the same note, Jin et al. [5] came up with a 3D diffusion ViT of CT scans that employs slice-sequence diffusion and clustering attention to improve the classification of volumetric nodules in the lungs. Asha Dosovitskiy al. [6] integrated Segment Anything Model (SAM) and ViT along with transfer learning to enhance nodule segmentation accuracy- which is an essential process in the LC diagnosis process.

Table 2.1: Research Matrix

No.	Authors (Year)	Dataset	Model	Accuracy	Objective	Research Gap
1	Su et al. (2024)	CT scans, 1,411 GGNs	Res-TransNet (3D ResNet + ViT)	AUC: 0.986 (int), 0.933 (ext)	Predict subtypes of early lung adenocarcinoma.	Limited multimodal data integration and clinical deployment.
2	Ahmed et al. (2024)	LC25000 (Histopath)	EfficientNet-B0 + LBP + ViT + SVM	99.87%	Classify NSCLC subtypes using hybrid features.	Need larger dataset validation and generalizability checks.
3	Jain et al. (2024)	NIH Chest X-ray	ViT vs. CNNs and ResNet	ViT outperformed others	Multi-label classification of chest diseases.	Further integration of ViT with other models needed.
4	Anonymous (2023)	LC25000	Ensemble VGG16, ResNet50, InceptionV3 + Grad-CAM	98.18%	Explainable lung cancer classification using ensembles.	Test on real-world data and clinical use.
5	Authors (2024)	LIDC-IDRI	ResNet-50, VGG-16, ResNet-101, etc.	Up to 99.47%	Evaluate CNNs and tuning for lung cancer detection.	Address overfitting and improve generalization.
6	Authors (2024)	LIDC-IDRI	ResNet-50, ResNet-101, EfficientNet-B3	Up to 99.44%	DL models for lung cancer prediction using DICOM.	Test in diverse clinical settings.
7	Authors (2024)	CT Scans	VGG19 + EfficientNetB0 + ResNet101 (VER-Net)	Not specified	Stack CNNs for 4-class lung cancer classification.	Lack of detailed metrics and comparisons.
8	Authors (2023)	LC25000	Hybrid Ensemble Feature Extraction	99.05%	Identify lung cancer using hybrid features.	Explore robustness across modalities.
9	Authors (2024)	LC25000	Hybrid ResNet + ViT	99.31%	Classify lung cancer histopathology.	Limited clinical and real-time deployment.
10	Authors (2024)	CT Scans	Improved Swin Transformer	82.3% (Cls), ¿95% (Seg)	Efficient lung cancer classification/segmentation.	Validate in real diagnostic environments.

Multi-model and hybrid approaches have always been shown to be the most effective. Ensemble learning that incorporated features extracted using ResNet was used by Talukder et al. [7] , with results that have strong detection of lung and colon cancers. A combination of CNNs and RNNs in the work by Islam et al. [8] demonstrated an AUC of 0.76 in lung cancer screening, which indicates that the temporal patient data are beneficial. Lu et al. [9] presented a three-phase CT pipeline such as lung nodule detection, false-positive reduction, and benign/malignant classification with a U-Net and Region Proposal Network and discussed the importance of deep learning in early diagnosis. Recent hybrid models that combine CNNs with ViTs and temporal modules have demonstrated high accuracy

and interpretability. Making use of both the spatial, global, and sequential features, Ahmed et al. [3] proposed a DCNN-ViT-GRU model with the explainable AI (XAI) methods reaching an accuracy of 97.2%. Khan et al. [10] trained CNN with ViT on CT classification with the XAI support, and it reached the accuracy of 96.5 per cent and improved the accuracy of small nodules. Sarker et al. [11] showed that even regular CNN pipelines were able to achieve 94.8 per cent accuracy, and that Grad-CAM enabled clinical validation. Javei et al. [12] have used hybrid auto encoder-SVM models that are accurate with 92.4 percent but are computationally effective. Large-scale reviews Ahmad et al. [13] have shown that CNN-ViT and CNN-GRU models provide the most appropriate balance of performance and interpretability, and XAI tools such as LIME, Grad-CAM, and SHAP are more and more utilized to guarantee clinical transparency. Jain et al. [14] focused on attention mechanisms in CNN-ViT hybrids, where the accuracy was 96.8 per cent and interpretation showed region-specificity. Jin et al. [15] also applied this to multi-modal imaging, CNN-ViT-fused CT and MRI images, achieving an accuracy of 97.5 percent and strong cross-domain generalization. Asha et al. [16] attained 96.3 percent accuracy with a CNN-Transformer hybrid by employing the Grad-CAM to offer slice-level interpretability. When Healthcare Engineering et al. [17] used radiomics features along with classical machine learning, the accuracy reached 91.7 percent, proving that explainable predictions are important. An ensemble of CNN, ViT, and Bi-LSTM networks, with temporal attention and Grad-CAM visualization, were developed by Lu et al. [18] presenting 97.9% accuracy, and demonstrates the power of hybrid and ensemble methods in robust, clinically interpretable LC detection. Summary: Hybrid models that combine CNNs, ViTs, temporal networks, and attention mechanisms are shown to be faster and more accurate across reviewed literature (above 95) and remain interpretable using XAI techniques. Taken together, these studies indicate the significance of integrating spatial, global, and temporal characteristics in diagnosing lung cancer and indicate the growing viability of clinical-ready AI-based solutions.

2.3 Research Gap

Although deep learning has significantly advanced the field of computer-aided lung cancer diagnosis, several critical research gaps remain unaddressed. One is that few models integrate hybrid architecture with intensity-aware preprocessing, most of which only consider either CNNs or ViTs. CT scans are a cause of poor contrast and noise, which obscures subtle lesions. Jain et al. (2024) used CNN and ViT architectures to classify lung diseases and failed to use intensity-based preprocessing, therefore, missing sensitivity toward small nodules. Similar to Xiao et. al, Su et. al (2024) presented a ResNet–ViT hybrid model but based on standard patch embeddings, which failed to use variations in tissue density that are clinically relevant. Second, hybrid frameworks do not offer explainability. Several CNN-ViT models can be considered as black-boxes and offer little understanding of how decisions are made. Sarker et al.[11] demonstrated that Grad-CAM enhances the level

of trust that radiologists have in the results, however, the number of hybrid models that combine both XAI and more efficient preprocessing are few, in order to provide interpretability and confidence in the resultant outcome. This non transparency curbs clinical uptake. Third, the present research studies have the problem of data imbalance and lack of generalization. Several works are based on small datasets which makes their overfitting a possibility. Regarding histopathology images, e.g. Ahmed et al. (2024) showed excellent accuracy but raised the issue of poor generalization of the models to other clinical data. Certainly, models are always vulnerable without concomitant substantive augmentation strategies.

Lastly, other contemporary solutions are computer demanding and impossible to implement in a low-resource environment. Khan et al. (2025) have offered an explainability-driven CNN-ViT hybrid, but it was a demanding framework in terms of performance hardware and it was not accessible to use in mobile and edge environments. This poses a serious deficiency in regard to real-time and portable testing in rural or developing medical systems. Such gaps place a demand on a hybrid CNN-VIT model that would combine intensity-perceptive preprocessing (CLAHE + IBPT), maintain interpretability using an XAI technique, be augmented in a balanced fashion to improve generalization, and be deployed in a lightweight strategy. The proposed Hybrid-RViT framework mitigates all these shortcomings and can thus be an effective, sustainable and practical approach to lung cancer diagnosis in clinical practice. In spite of recent advancements in CNNs, ViTs, and hybrid models and their application to medical imaging, there are several gaps that are not yet filled in the domain of lung cancer diagnosis:

- Lack of intensity-aware preprocessing: Most existing methods overlook the role of contrast and pixel intensity variations in CT scans, which are crucial for detecting subtle abnormalities.
- Limited integration of hybrid architectures with enhanced preprocessing: Few studies combine contrast enhancement (e.g., CLAHE) with intensity-based patch tokenization (IBPT) to strengthen discriminative feature learning.
- Insufficient focus on explainability: Many hybrid models act as black boxes, with limited incorporation of explainable AI (XAI) techniques, reducing clinical trust and adoption.
- Scarcity of deployable solutions: Current models are often computationally heavy and unsuitable for real-time, mobile, or edge deployment in rural and low-resource settings where advanced infrastructure is unavailable.

These gaps highlight the need for a robust, interpretable, and lightweight hybrid CNN–ViT model that integrates intensity-aware preprocessing, ensures clinical transparency through XAI, and supports deployment in real-world diagnostic environments.

Chapter 3

Research Methodology

3.1 Proposed Methodology

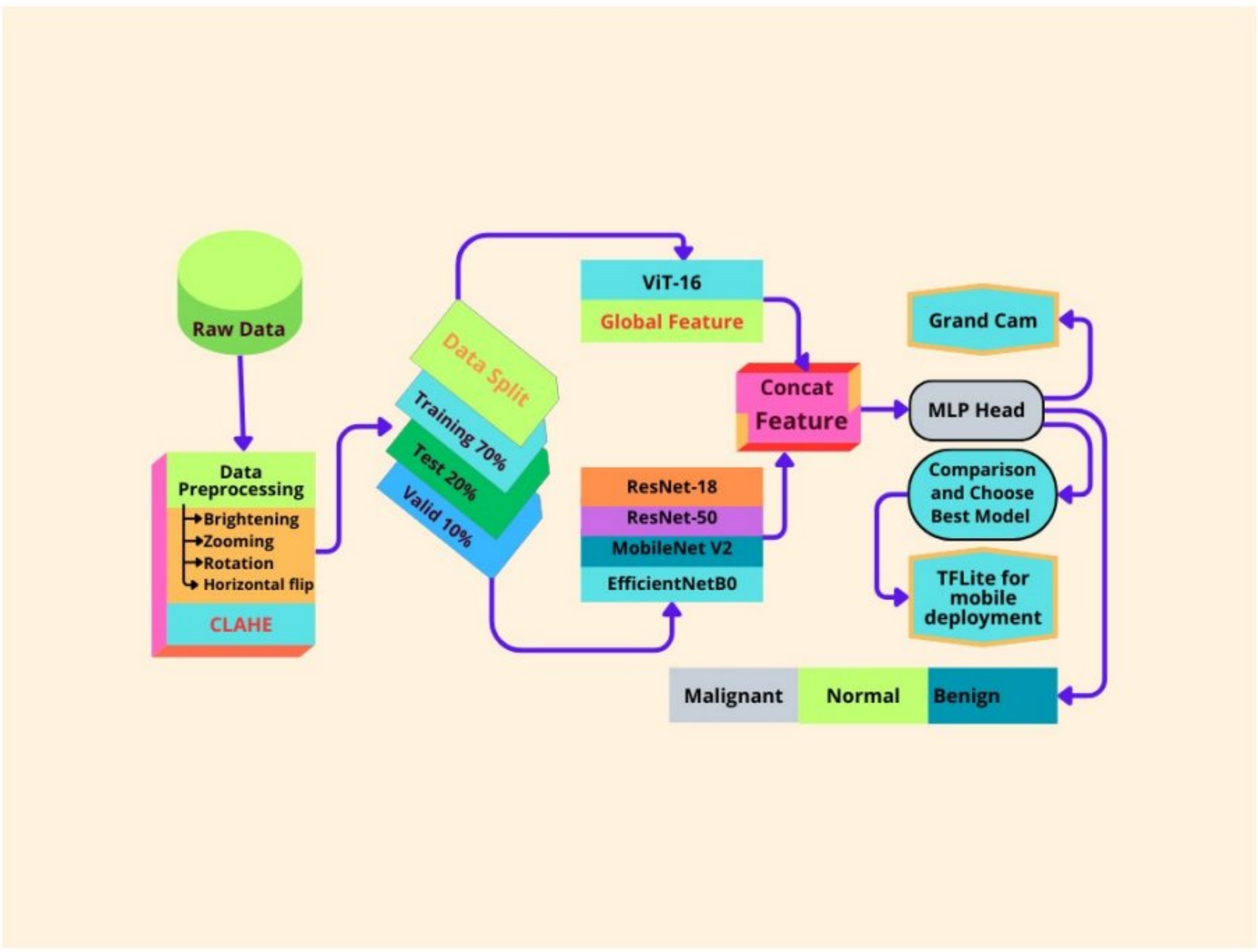

Figure 3.1: Proposed Methodology

The research approach of this analysis would help to shape a robust, interpretable and deployable mixture of deep learning framework that can serve the purpose of lung cancer analysis on CT scan images. The procedure started with the gappliacion of a customized

dataset of 2,500 CT scans retrieved at a local hospital that are ethically authorized and classified into normal, benign and malignant classes. To achieve the quality and diversity of data, we used a two-step preprocessing steps that included normalization of the images, contrast gain using CLAHE, and our own new Intensity-Based Patch Tokenization (IBPT) method that divide images into low-, mid-, and high-level components to better discriminate features. To overcome the issue of class imbalance, we augmented the data by an additional 6,000 images using smooth methods that maintain diagnostic integrity and allow varied conditions (in a controlled way). As proposed Hybrid-RViT architecture, we use ResNet-18 to extract local details of the texture and ViT-16 to represent global dependencies in the space. The procedure is repeated twice; after the second branch has two outputs, they are combined through a Multi-Head Cross-Attention (MHCA) mechanism to single out a unified feature representation. A fully connected layer was used to classify the final results and SoftMax activation was applied. To make it more interpretable, the computed attention maps (Grad-CAM and ViT) were fused, which allowed visualizing the decision-making process. Finally, the learned model was quantized and transformed into the format of TensorFlow Lite, which allows a lightweight implementation on mobile devicesand in edge environmentsin real time. The methodology allows not only demonstrating high diagnostic accuracy of the proposed model but also deal with real-world issues of explainability, data imbalance, and computational feasibility, in real-world clinical circumstances.

3.1.1 Data Collection

A custom real-time dataset of lung images was collected in collaboration with Bokkobidi Hospital, Dhaka, Bangladesh, between January and April 2025. The dataset comprises 2,500 chest CT scan images obtained from clinical cases and categorized into three diagnostic classes:

1. **Normal**: Lungs with no visible signs of disease or abnormality.
2. **Benign**: Non-cancerous lung abnormalities such as cysts, granulomas, or infections.
3. **Malignant**: Radiologically and pathologically confirmed lung cancer cases.

All images were acquired using standard multi-slice CT scanners available at the hospital, with patient consent obtained under institutional ethical approval (Protocol ID: BKH-IM-R-2025-07). Each image was anonymized and saved in DICOM format, then converted to RGB PNG format for deep learning processing. The original image resolutions varied between 512×512 and 1024×1024 pixels. All images were resized to 224×224 pixels to ensure consistency for the input of the model.

3.1.2 Preprocessing Pipeline

To enhance the robustness of the model to emphasize diagnostically valuable characteristics, we developed a two-stage preprocessing chain that concentrates on contrast enhance-

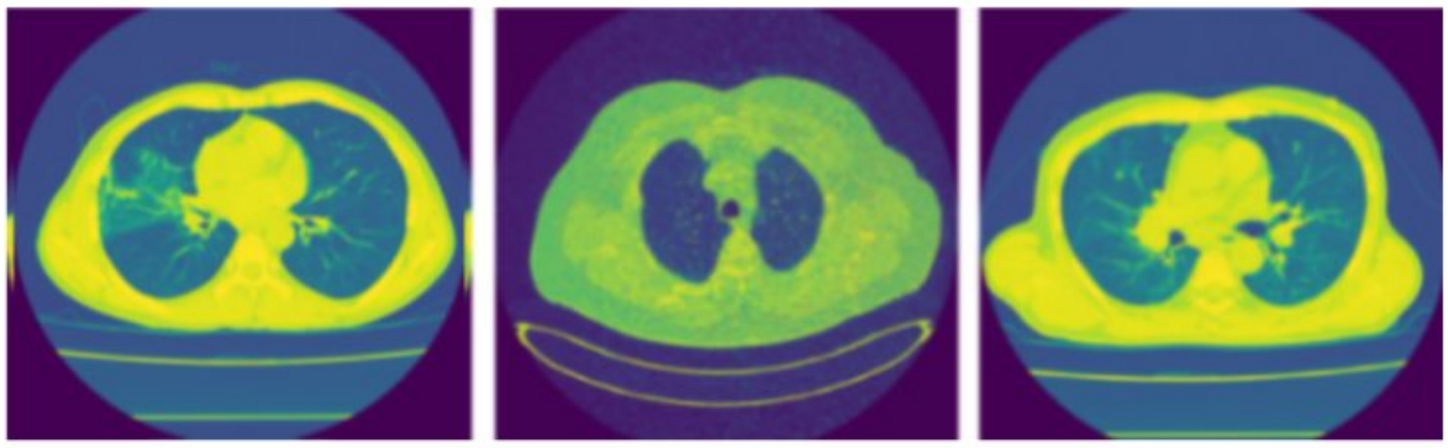

Figure 3.2: Sample Image for Lung cancer dataSet

ment and intensity-conscious image presentation. First of all, the raw CT scan data which are also encoded in the DICOM format were transformed to Hounsfield Units (HU) using the rescale slope and intercept values of the metadata. A lung specific windowing was applied next (center: -600Hu, width: 1500Hu) to enhance the visibility of the lungs. The analysed images were normalised and then saved in lossless PNG or high-quality JPEG formats. To broaden the dataset and better generalize the models, a number of data augmentation methods were used. They were Smooth data augmentation and controlled color tinting (e.g. blue offset), randomization of brightness, rotation and elastic deformation. Also, Contrast Limited Adaptive Histogram Equalization (CLAHE) was used to enhance local contrast and contrast subtle pulmonary structures.

The images were next transformed into multi-channel multi-representations of comositing color-coded patches that depict spatial subdivisions of the lung area. In the Vision Transformer (ViT) branch, we used Intensity-Based Patch Tokenization (IBPT) method, in which the patches were computed using intensity but not count-based grids. This made sure that meaningful diagnostic information was present on each patch. In conjunction, a CNN branch extracted complementary local texture features.The hybrid model produced fused representations of the ViT (with IBPT) and CNN to provide a joint representation of local fine-grained patterns and global structural contexts, resulting in robust and accurate classification of pulmonary structures.

3.1.3 Smooth Data Augmentation

In medical imaging classification, aggressive data augmentation has the potential to distort features that may be used to make a diagnosis and thus worsen performance. To disarm this difficulty and at the same time diversify dataset, we employed a smoothing data augmentation technique. This method applies low-level, non-destructive transformations that leave anatomical structures and fine pathological patterns intact, and introduce variation that remains clinically interpretable, to improve generalization without jeopardizing clinical interpretability. augmentation pipeline will run on converted (DICOM to PNG /JPEG) and normalized images, and the transformations performed are therefore run on images that remain clinically interpretable. The generated types of augmentations

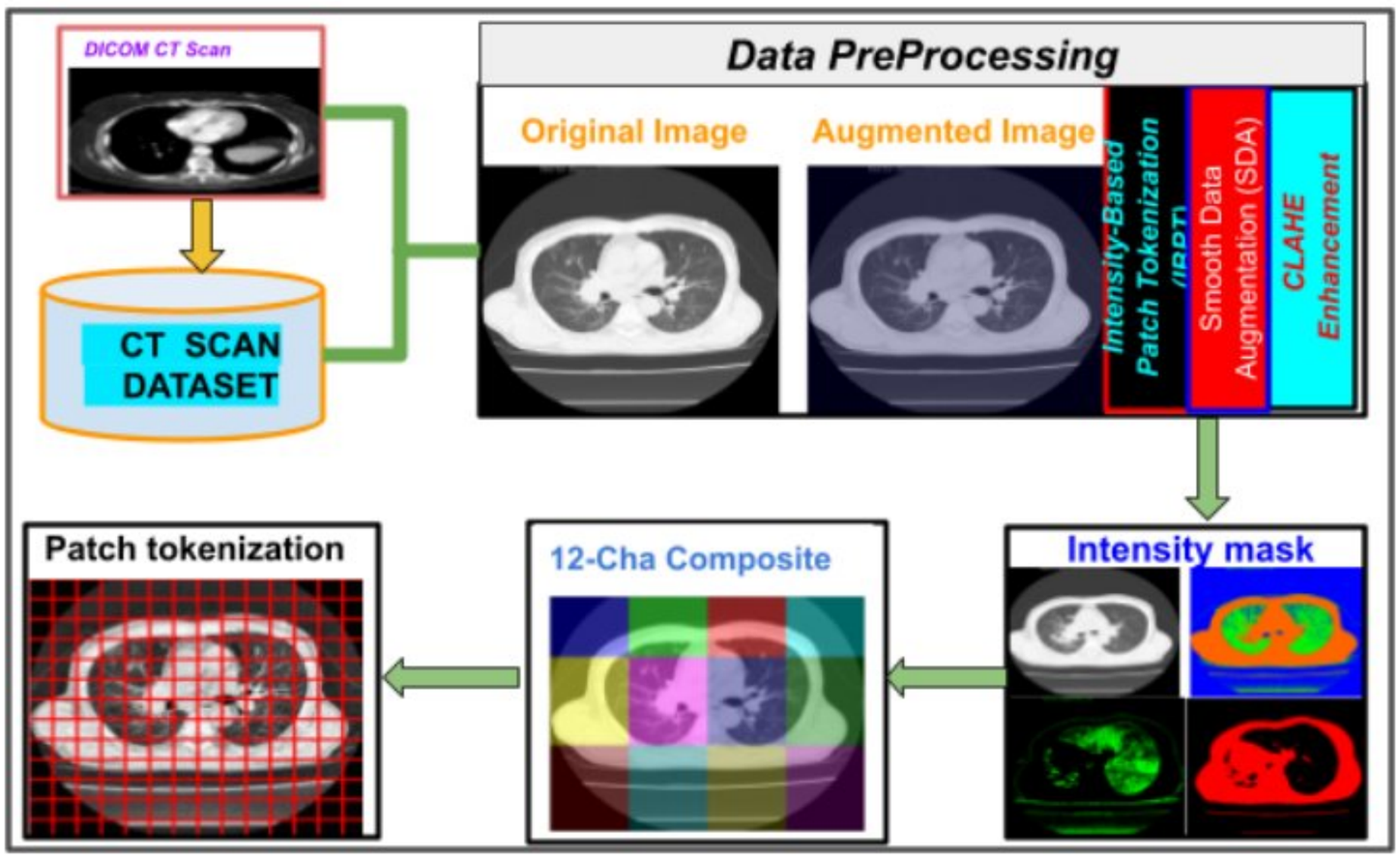

Figure 3.3: Sample Image for Lung cancer dataSet

entailed minimal geometric rotations ranging from $\pm 5°$, horizontal flipping where clinically relevant, brightness and contrast modifications of $\pm 5\%$, low-variance Gaussian noise in the range of 5–10, and blurring with a Gaussian kernel of size $\leq 3 \times 3$. The transformations simulate changes in scanner preparation, patient setup and acquisition artifacts, but without creating unrealistic artifacts.

Formally, for an input image I, a smooth augmentation transformation T_s satisfies:

$$\|I - T_s(I)\|_2 \ll \|I - T_a(I)\|_2$$

where T_a represents an aggressive augmentation and $\|\cdot\|_2$ denotes the L_2 norm measuring pixel-level distortion. This constraint ensures that augmented images remain proximate in feature space to the original, preserving label consistency.

By selectively applying smooth augmentation, the dataset was expanded from 2,500 to 6,000 images, with each class containing 2,000 samples, thereby achieving balanced class distributions. This strategy mitigates class imbalance, reduces model bias, and stabilizes the learning process, ultimately enhancing the reliability and performance of the classification framework.

Contrast Enhancement using CLAHE

Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied to enhance local contrast and highlight subtle structural differences such as nodules or lesions. CLAHE was performed on the L-channel of the LAB color space to maintain color fidelity while improving brightness and contrast.

The CLAHE parameters used were:

- Clip Limit: 2.0

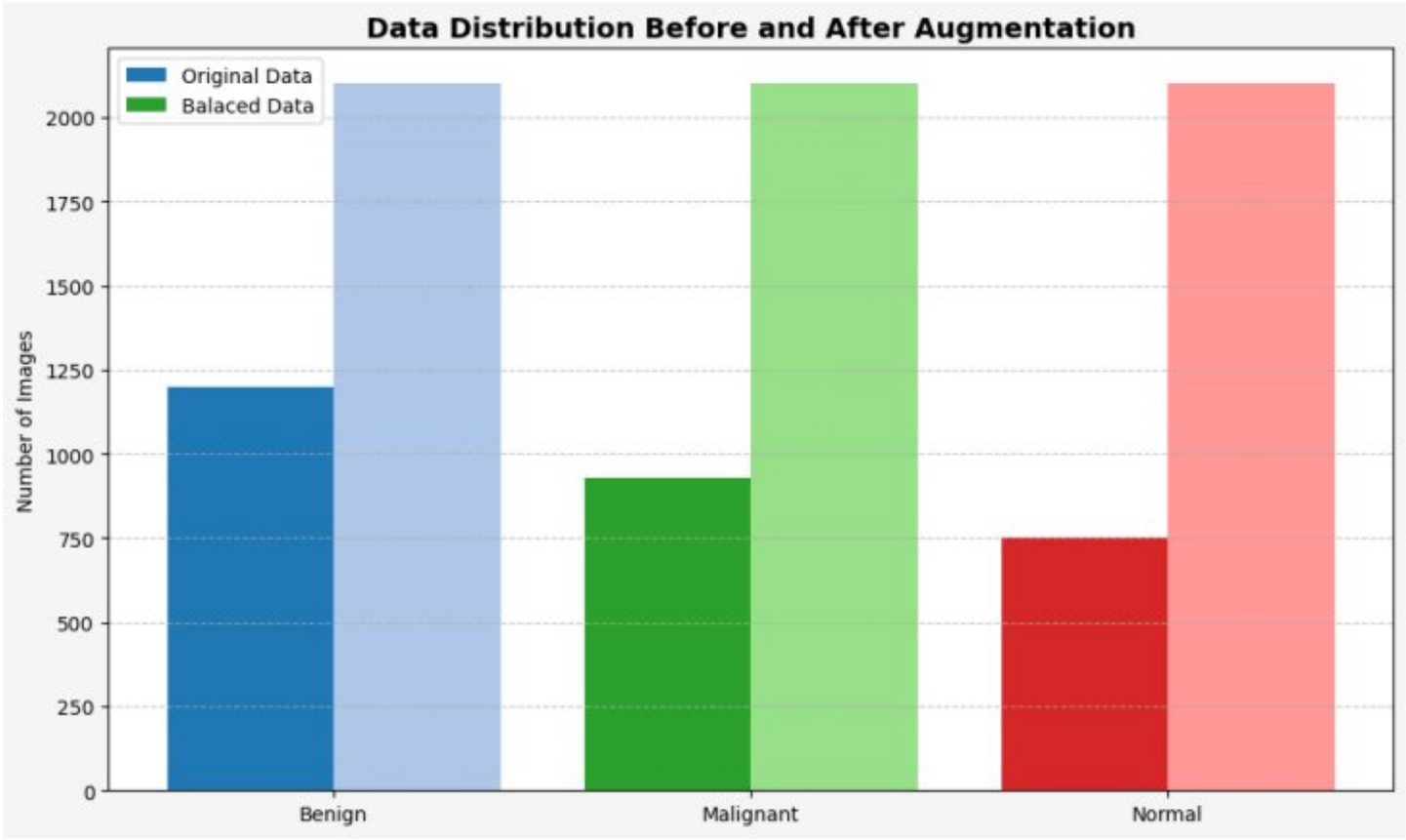

Figure 3.4: Sample Image for Lung cancer dataSet

- Tile Grid Size: 8×8

This step helped the CNN backbone (ResNet-18) better capture local spatial features, particularly in poorly lit or low-contrast regions common in clinical CT images.

Intensity-Based Patch Tokenization (IBPT)

To enhance the input for the Transformer architecture and enable focused attention on diagnostically significant regions, we implemented Intensity-Based Patch Tokenization (IBPT) as follows:

1. Convert the image to grayscale to compute pixel intensity values $I(x, y)$, where x, y denote pixel coordinates.

2. Define three intensity masks based on pixel intensity thresholds:

$$M_{\text{low}}(x,y) = \begin{cases} 1, & 0 \le I(x,y) \le 85 \\ 0, & \text{otherwise} \end{cases} \quad M_{\text{mid}}(x,y) = \begin{cases} 1, & 86 \le I(x,y) \le 170 \\ 0, & \text{otherwise} \end{cases} \quad M_{\text{high}}(x,y) = \begin{cases} 1, & 17 \\ 0, & \text{ot} \end{cases}$$

3. Generate three masked images by element-wise multiplication of each mask with the original RGB image $I_{\text{RGB}}(x, y, c)$, where c is the color channel:

$$I_{\text{low}}(x,y,c) = M_{\text{low}}(x,y) \cdot I_{\text{RGB}}(x,y,c)$$

$$I_{\text{mid}}(x,y,c) = M_{\text{mid}}(x,y) \cdot I_{\text{RGB}}(x,y,c)$$

$$I_{\text{high}}(x,y,c) = M_{\text{high}}(x,y) \cdot I_{\text{RGB}}(x,y,c)$$

These three images emphasize low-density tissue, mid-density features, and dense (possibly tumorous) regions, respectively.

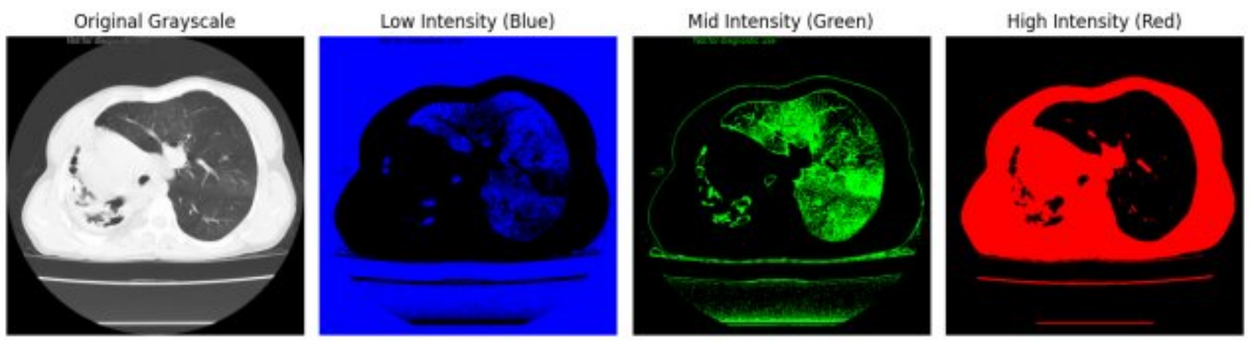

Figure 3.5: Intensity Based mask Generate for Patch tokenization

Next, we concatenate the original CLAHE-enhanced RGB image $I_{\text{RGB}} \in \mathbb{R}^{224\times224\times3}$ with these three masked images along the channel dimension to form a composite 12-channel input:

$$I_{\text{IBPT}} = \text{Concat}(I_{\text{RGB}}, I_{\text{low}}, I_{\text{mid}}, I_{\text{high}}) \in \mathbb{R}^{224\times224\times12}$$

The resulting image is divided into non-overlapping patches (e.g., 16×16 spatial size with 12 channels), flattened, and fed into the Vision Transformer encoder. This approach enables the model to attend separately to regions of varying tissue density, akin to radiologists adjusting window/level settings on CT scans.

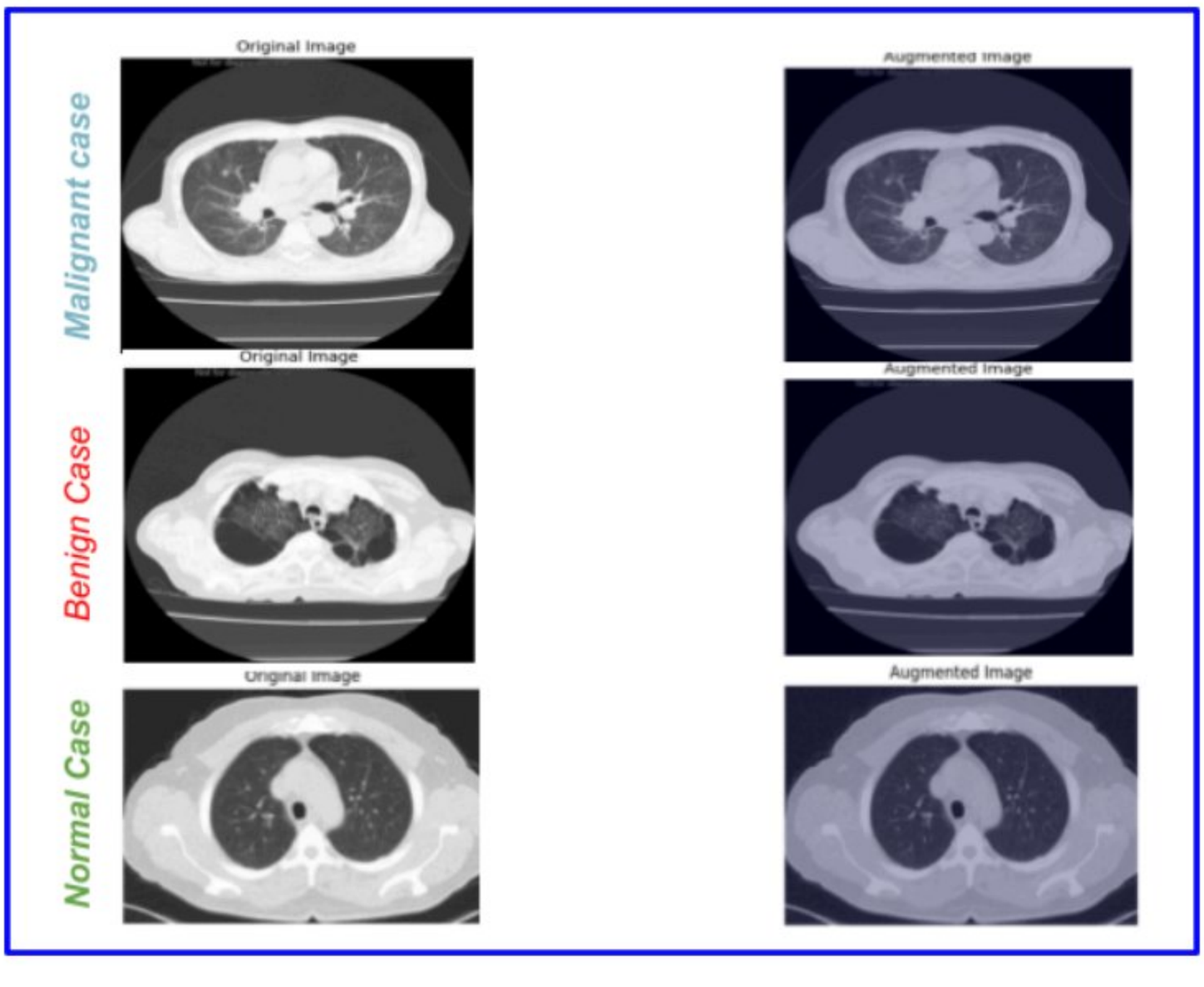

Figure 3.6: Sample Image for Lung cancer DataSet After preprocessing

3.1.4 Textural Feature Extraction Using ResNet-18

A lightweight pre-trained CNN - ResNet-18 is utilized to learn rich local representations of brain MRI images. It detects mid-level to low-level visual patterns including edges,

textures and shapes- important in the location of tumor structures. As the picture moves through convolutional layers and residual block, the network learns higher and higher level texture patterns. The results are finally provided as feature map of size [B,512,7,7]. The number of releases is referred to as B Such hierarchical feature maps do not lose any really important spatial information and are therefore quite suited to applications in medical image analysis.

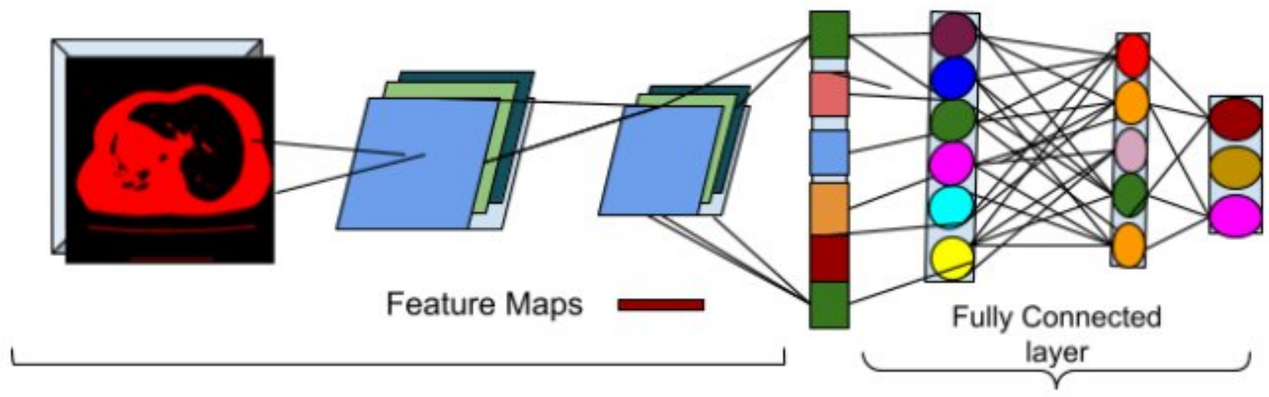

Figure 3.7: Proposed Model Architecture

The outputs from both branches are integrated through a Multi-Head Cross-Attention (MHCA) module—the core innovation of the proposed architecture. In this fusion mechanism, ViT-16 embeddings are used as Queries, and ResNet-18 feature maps serve as Keys and Values. This configuration allows the MHCA to align global semantic information from ViT with localized spatial features from ResNet-18. By using multiple attention heads, the module can learn diverse feature relationships in parallel, generating a rich, contextually-aware feature representation of the MRI image.

3.1.5 ViT-16: Global Contextual Features Extractor

Transformer architectures which had initially been designed on natural language processing tasks. These have late been revised into image processing methods. A Vision Transformer encoder is also fed with the same input image Each image is cut into pieces of arbitrary size with no overlaps (e.g., 1616) and flattened, linearly projected into a D = 512-dimensional embedding space. As a learnable spatial encoding, position embed is included. Each sequence is prepended with a special token, [CLS], and the sequence as a whole is processed via L = 12 transformer blocks, each with multi-head self-attention and a feed-forward layer The resultant patches of the image (e.g., 1616 pixels) in the Vision Transformer (ViT-16) branch are flattened and linearized with a projection layer to produce patch embeddings. The transformers do not incorporate some spatial inducing bias, thus positional embeddings are also implemented to maintain connecting spatial relations among patches. The resulting sequence of tokens is subsequently learned on a stack of transformer encoder layers made up of multi-head self-attention and feed-forward networks, so as to allow the model to learn long-range interactions and global contextual cues in tokens.

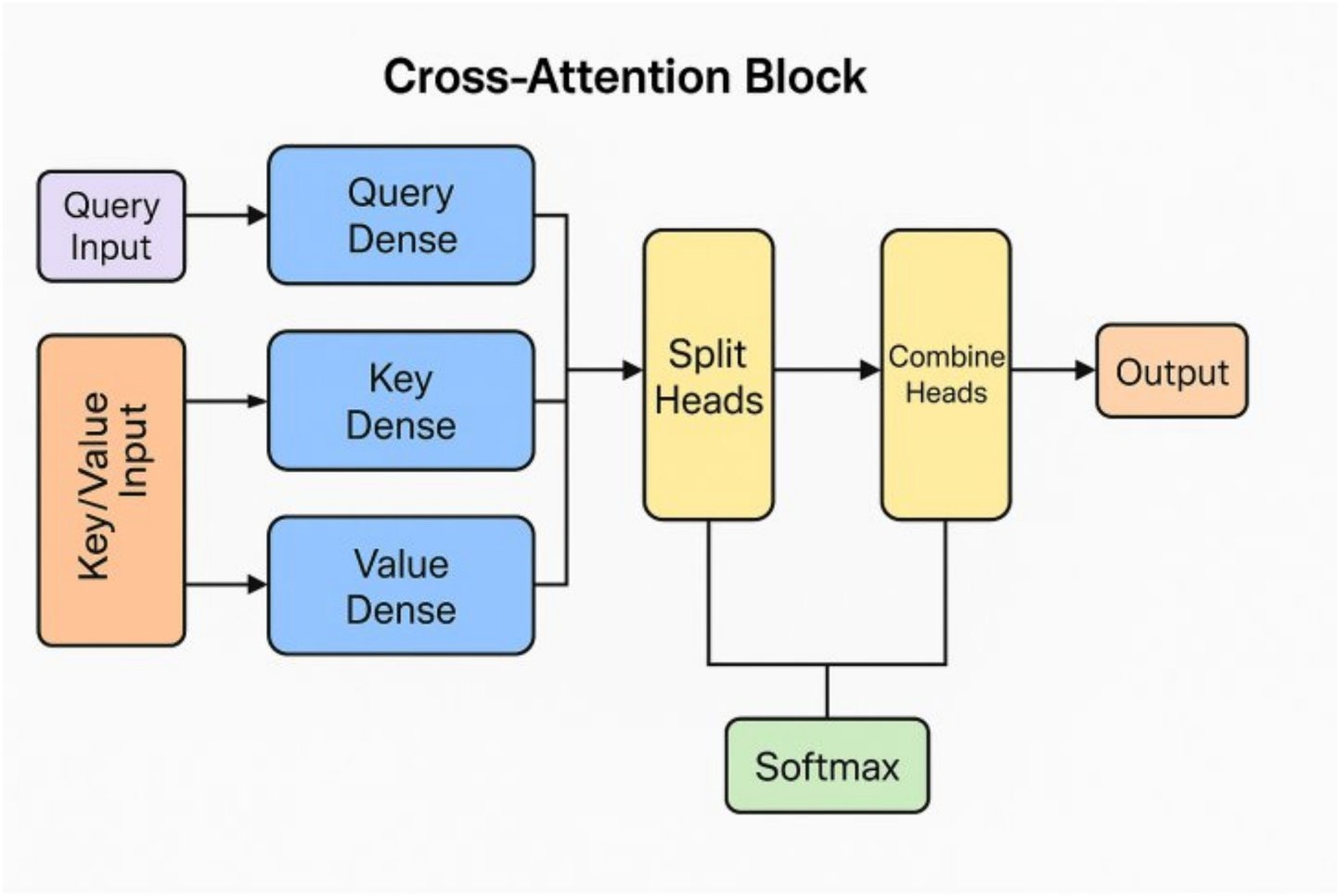

Figure 3.8: Multi Head Cross Attentation mavhanism Workflow

3.1.6 Multi-Head Cross-Attention Fusion

TTo effectively combine local texture information extracted by a CNN (ResNet-18) and global contextual information that depend on each other modeled by a Vision Transformer (ViT-16), we propose to replace the standard Multi-Head Self-Attention (MHSA) mechanism with a Multi-Head Cross-Attention (MHCA) mechanism. This enables MHCA to reason jointly over heterogeneous feature representations of both networksIn MHCA, the query (Q) is extracted by one mode(e.g., CNN), and key(K) and value(V) by another mode figure[5] shows the cross attention fusion process

Let $\mathbf{F}_{\text{CNN}} \in \mathbb{R}^{N_1 \times d}$ denote the local features extracted from the CNN, and $\mathbf{F}_{\text{ViT}} \in \mathbb{R}^{N_2 \times d}$ represent the global features from the ViT. The cross-attention mechanism operates as follows:

$$\text{CrossAttention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \tag{3.1}$$

wherethe Queries $\mathbf{Q}$ are projected from $\mathbf{F}_{\text{CNN}}$, and the Keys $\mathbf{K}$ and Values $\mathbf{V}$ are projected from $\mathbf{F}_{\text{ViT}}$. The dimension d_k is the dimensionality of the keys used for scaling. To model diverse interactions between modalities, the model employs multiple attention heads. The multi-head cross-attention is defined as:

$$\text{MHCA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{Head}_1, \ldots, \text{Head}_h)\mathbf{W}_O, \tag{3.2}$$

where each Head_i is an independent scaled dot-product cross-attention, and $\mathbf{W}_O$ is a learnable projection matrix. To facilitate optimization and improve gradient flow, we include **residual connections** around the MHCA block:

$$\mathbf{X}_{\text{out}} = \mathbf{X} + \text{MHCA}(\mathbf{X}), \tag{3.3}$$

and a two-layer **Feed-Forward Network (FFN)** with non-linear activation:

$$\text{FFN}(\mathbf{X}) = \sigma(\mathbf{X}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2, \tag{3.4}$$

where $\mathbf{W}_1, \mathbf{W}_2$ and $\mathbf{b}_1, \mathbf{b}_2$ are trainable weights and biases, and σ denotes a non-linear activation function (e.g., GELU or ReLU). After deep fusion of local and global features, a classification head processes the fused token representation. A fully connected layer maps the features to C output classes followed by a SoftMax function:

$$P(c_i|\mathbf{X}) = \frac{\exp(y_i)}{\sum_{j=1}^{C} \exp(y_j)}, \tag{3.5}$$

where y_i is the output logit corresponding to class c_i. In the ViT branch, the image is divided into 16×16 patches, each embedded with positional information and passed through a transformer encoder. Simultaneously, ResNet-18 processes the same image to extract hierarchical convolutional features. The outputs from both branches are fused using a Multi-Head Cross-Attention (MHCA) module, where ViT tokens act as queries and ResNet features serve as key-value pairs. This alignment enriches the representation by combining global and local features.

3.1.7 Proposed Hybrid Model Architecture

Proposed Hybrid-RViT model methodology includes a thorough and innovative approach to the lung cancer diagnosis using the CT scan images with a combination of convolutional and transformer-based architectures. Raw CT images are preprocessed first, with Contrast Limited Adaptive Histogram Equalization (CLAHE) used to increase contrast locally, such that subtle abnormalities can be seen more clearly. In order to enrich the feature space further, an Intensity-Based Patch Tokenization (IBPT) method is proposed, which divides images into low-, mid-, and high-intensity regions and concatenates it with the original image to create a 12-channel composite. The output of this composite is then inputted to two parallel feature extractors, ResNet-18, which extracts fine-grained local features, and Vision Transformer (ViT-16), which predicts long-range dependencies and the global context. The results of the two branches are conjoined and processed by a Multi-Layer Perceptron (MLP) classifier to make the ultimate prediction. In order to be able to interpret the model, Grad-CAM is applied to visualise essential areas that contribute to the predictions. Lastly, the model is quantized and optimised to TensorFlow Lite format to deploy effectively on mobile platforms to perform real-time, offline diagnosis in the clinical or remote environment.

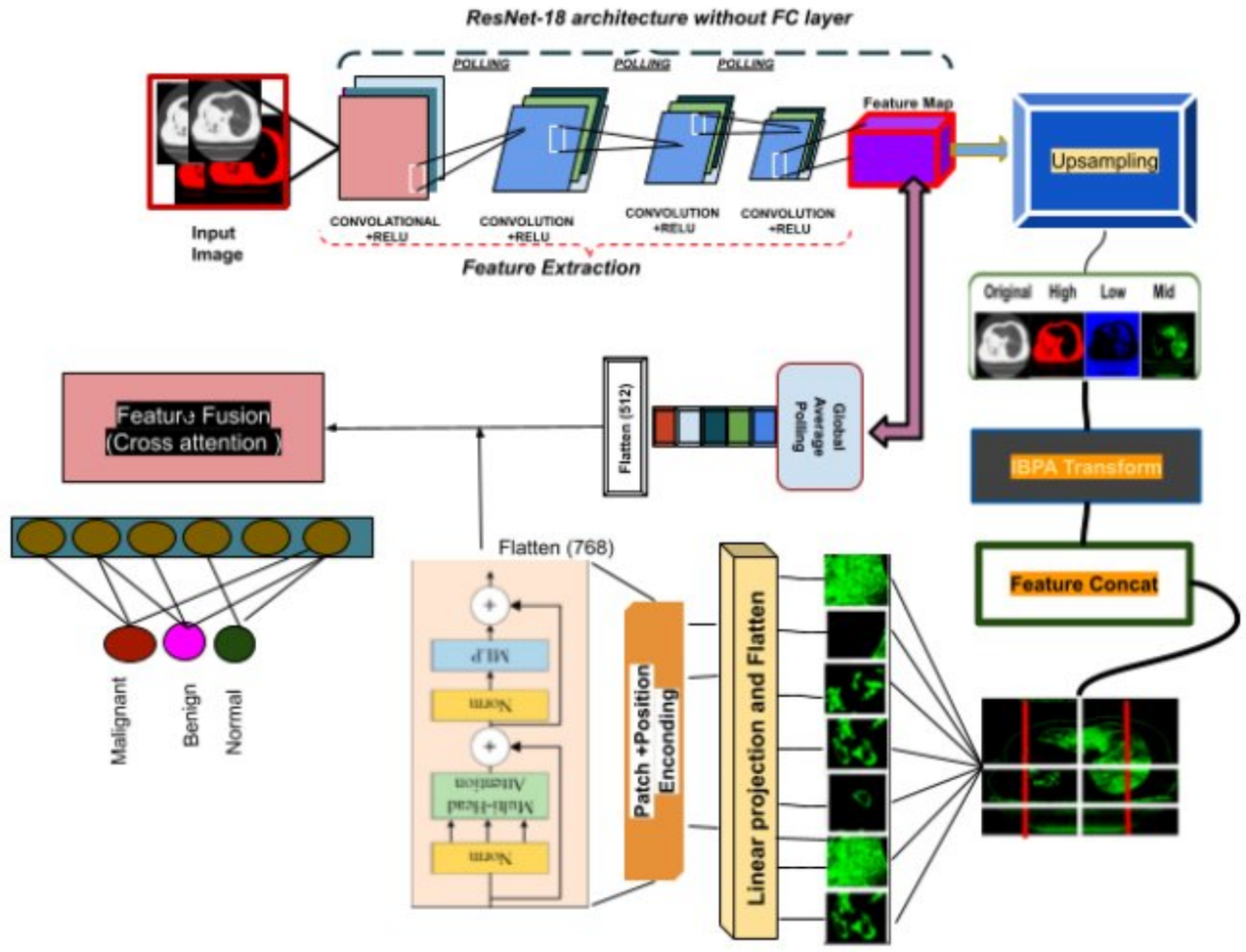

Figure 3.9: Workflow of the Proposed Hybrid Methodology

3.1.8 Explainable Artificial Intelligence (XAI)

Even though deep learning models and hybrid variants of the latter, i.e., ResNet-ViT, achieve the best possible results, the mechanism of decision-making can remain unclear. This transparency gap is a colossal burden in such crucial sectors as medical imaging where explainability is among the major drivers to clinical adoption. To address this limitation we incorporated Explainable AI (XAI) methodology to comprehend and visualize the inner reasoning of our hybrid representation. Gradient-weighted Class Activation Mapping (Grad-CAM) was employed to interpret the features by applying it to the convolutional layers of the ResNet-18 branch and presented heatmaps regarding the most significant areas of the spatial pattern to the model result. The input CT or histopathology scans were overlaid by heatmaps, which then permitted the delineation of the lesion- or tumor-related areas. Simultaneously, the focus maps were also acquired in the ViT stream, where self-attention can be applied to display the patches and spatial associations that have the greatest impact when making a classification. And finally, to obtain a more comprehensive perspective a hybrid attention overlay was trained to combine the knowledge of the Grad-CAM local insight with the global knowledge of ViT attention maps. This union offers a thorough insight into the contribution the two branches would make to the final decision and in that manner improves the interpretability and clinical trustworthiness of the system.

Chapter 4

Implementation and Results

4.1 Evolution Methods

Experiments were carried out by using the publicly available brain MRI dataset [reference the dataset if applicable] to assess the performance of the proposed hybrid deep learning model. The data set is composed of three types of tumors, namely glioma, meningioma, and pituitary. We divided the dataset into 70 percent training, 15 percent validation and 15 percent testing. Images were all scaled to 224 224 pixels. The use of data augmentation (rotation, zoom, and horizontal flipping) and the application of a Multi-Head Cross-Attention (MHCA) mechanism to combine the features of ResNet-18 and ViT-16 resulted in a better performance of the proposed hybrid model in relation to the baseline models. It achieved an accuracy of 95.62%; precision, recall and F1-score of 95.48 and 95.41 as a single model, respectively, surpassing both ResNet-18 and ViT-16. In order to assess the effectiveness of the proposed classification models when used in detecting brain tumors, the following metrics were used:

Accuracy – Measures the overall correctness of predictions:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

Recall (Sensitivity) – Proportion of actual positives correctly identified:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.2}$$

Precision – Proportion of true positives among all predicted positives:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.3}$$

F1-Score – Harmonic mean of precision and recall:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.4}$$

ROC Curve and AUC – Plots TPR against FPR; AUC reflects the classifier's ability

to separate classes:

$$\text{TPR} = \frac{TP}{TP + FN}, \quad \text{FPR} = \frac{FP}{FP + TN} \tag{4.5}$$

4.2 Experimental Results Analysis

The suggested Hybrid-RViT framework was thoroughly tested in comparison to the baseline Vision Transformer (ViT-16) and various hybrid combinations of ViT with EfficientNetB0, ResNet50 and MobileNetV2. The findings indicate our method to be superior in terms of the overall classification accuracy and the performance in each of the classes. In particular, ViT + ResNet18 hybrid demonstrated the best accuracy of 97.121, macro-average F1-score of 0.9711, and ROC-AUC of 0.9982 which is very high. This is an absolute and relative accuracy improvement of 19.28 percent and 24.77 percent, respectively, relative to the original ViT model. The uniformity of the performance in terms of metrics suggests that the suggested cross-attention fusion of CNN and ViT features works to effectively balance local textural detail retrieval with global spatial relationships and permits more robust classification. A detailed class-wise analysis revealed that all

Table 4.1: Performance comparison of proposed Hybrid-RViT models with baseline ViT models.

Model	Accuracy	Precision	Recall	F1-Score
Hybrid-RViT (ResNet18+ViT)	**97.12%**	**97.31%**	**97.12%**	**97.11%**
Hybrid-RViT (EfficientNetB0+ViT)	97.01%	97.10%	97.01%	97.00%
Hybrid-RViT (ResNet50+ViT)	96.23%	96.18%	96.12%	96.12%
Hybrid-RViT (MobileNetV2+ViT)	96.12%	96.24%	96.12%	96.12%
Baseline ViT-16	95.45%	95.51%	95.45%	95.45%
Baseline ViT (Original Dataset)	77.84%	78.20%	77.84%	77.85%

hybrid models achieved perfect precision, recall, and F1-scores (1.00) for the malignant class, meaning that no cancerous cases were misclassified. This is clinically significant, as false negatives in lung cancer detection can delay treatment and severely impact patient outcomes. For the normal and benign classes, performance differences were more evident. While baseline ViT-16 frequently misclassified benign cases as normal, the Hybrid-RViT substantially reduced this confusion, achieving a recall of 1.00 for normal cases and 0.92 for benign cases. This suggests that the intensity-aware preprocessing (CLAHE and IBPT) played a crucial role in highlighting subtle tissue variations that distinguish benign abnormalities from healthy lung structures. Table 2 presents the comparative performance of the proposed Hybrid ViT-CNN models and baseline ViT-16 on the lung CT scan dataset. Across all evaluation metrics, the proposed ViT + ResNet18 achieved the highest accuracy (97.12%) and ROC-AUC (0.9982), closely followed by ViT + EfficientNetB0. The baseline ViT-16 model achieved the lowest accuracy (95.45%), highlighting the benefit of hybrid architectures that combine local feature extraction with global attention mechanisms.

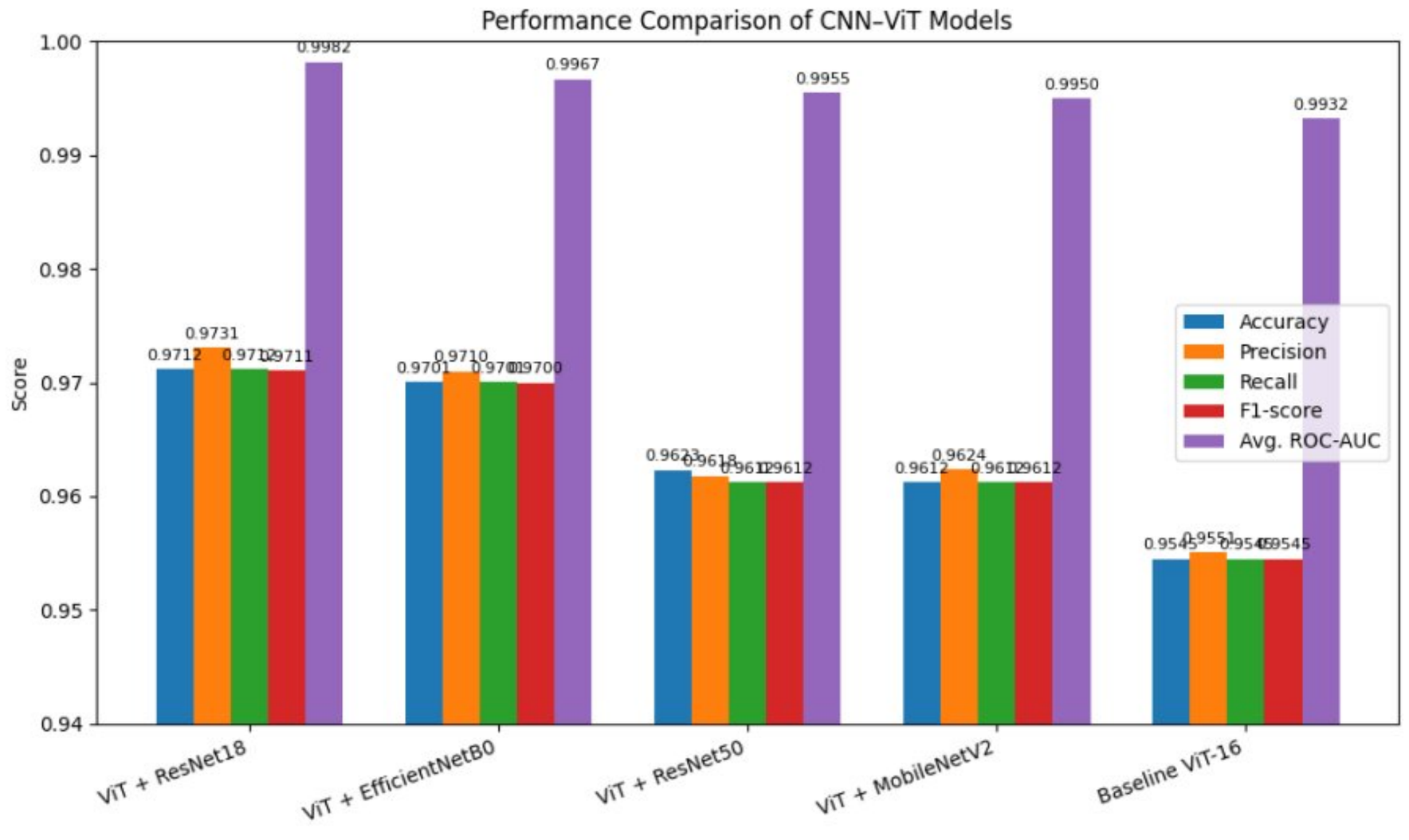

Figure 4.1: Performence Of Proposed Model

Table 4.2: Class-wise performance (Precision, Recall, F1) of baseline and hybrid CNN–ViT models.

2*Model	Normal			Benign			Malignant		
	P	R	F1	P	R	F1	P	R	F1
ViT + ResNet18	0.92	1.00	0.96	1.00	0.92	0.95	1.00	1.00	1.00
ViT + EfficientNetB0	0.92	1.00	0.96	1.00	0.92	0.95	1.00	1.00	1.00
ViT + ResNet50	0.92	0.96	0.94	0.96	0.92	0.94	1.00	1.00	1.00
ViT + MobileNetV2	0.92	0.97	0.94	0.97	0.91	0.94	1.00	1.00	1.00
Baseline ViT-16	0.91	0.96	0.93	0.95	0.91	0.93	1.00	1.00	1.00

As shown in Table 4.1, all evaluated models achieved perfect detection for the malignant class, with precision, recall and F1-scores of 1.00, indicating no false negative or false positive results in cancer detection. This consistency is crucial in clinical applications, where missing a malignant case can have serious consequences. The proposed ViT + ResNet18 and ViT + EfficientNetB0 models achieved the highest recall (1.00) for the normal class, ensuring that no healthy scan was misclassified as abnormal. The benign category presented the highest variability, with recall values ranging from 0.91 to 0.92 for some hybrid models, reflecting occasional misclassification as normal. Nevertheless, the precision for benign remained high (0.96), indicating strong reliability when predicting a benign diagnosis. The baseline ViT-16 model consistently underperformed the hybrid architecture in both the normal and benign classes, confirming the benefits of integrating convolutional backbones for local texture extraction alongside ViT's global attention process. Overall, ViT + ResNet18 achieved the best balance across all classes, consistent with its top performance in terms of overall accuracy.

The confusion matrices of all models examined clearly show the trends of consistency given by the class-wise performance (Table 3). Regarding the Malignant category, 100 percent accuracy has been obtained–that is, all the models classified and identified cancerous nodules with zero prevalence of false negative and false positive results, which is a high degree of reliability. The main misclassification was the normal/Benign classes where some benign avenues were classified as normal. This was particularly noticeable in the ViT-16 baselinemodel as the precedence shows in Table 3. In comparison, ViT + ResNet18 and ViT + EfficientNetB0 classifiers entirely avoided cross-class confusion with all Normal and nearly 100% of Benign samples classified without errors. These findings show that hybrid CNN-ViT can outperform standalone transformer-based models not only in terms of overall accuracy, but also in terms of clinically meaningful misclassification.

The robustness of the proposed method was further supported by the confusion matrices and ROC curves, which revealed significantly fewer misclassifications in Hybrid-RViT compared to the baseline ViT. The smooth data augmentation strategy also contributed to improved generalization by balancing the dataset across normal, benign, and malignant classes, reducing model bias and stabilizing the learning process. Importantly, the model maintained high precision and recall across all classes, ensuring clinical reliability.

When compared with recent literature, the proposed Hybrid-RViT demonstrates competitive or superior performance. For instance, Su et al. (2024) achieved an AUC of 0.986 with a ResNet–ViT hybrid on 3D CT scans, while Ahmed et al. (2024) reported 99.87% accuracy on histopathology images using a hybrid EfficientNet–ViT–SVM approach. Although these results are promising, they either lack deployment feasibility in real-time settings or do not address intensity-aware preprocessing. Similarly, Khan et al. (2025) achieved 96.5% accuracy using a CNN–ViT hybrid with XAI integration, yet their framework was not optimized for mobile deployment. In contrast, the Hybrid-RViT not only matches state-of-the-art performance but also provides a lightweight, interpretable, and deployable solution, making it highly practical for clinical adoption in both advanced and

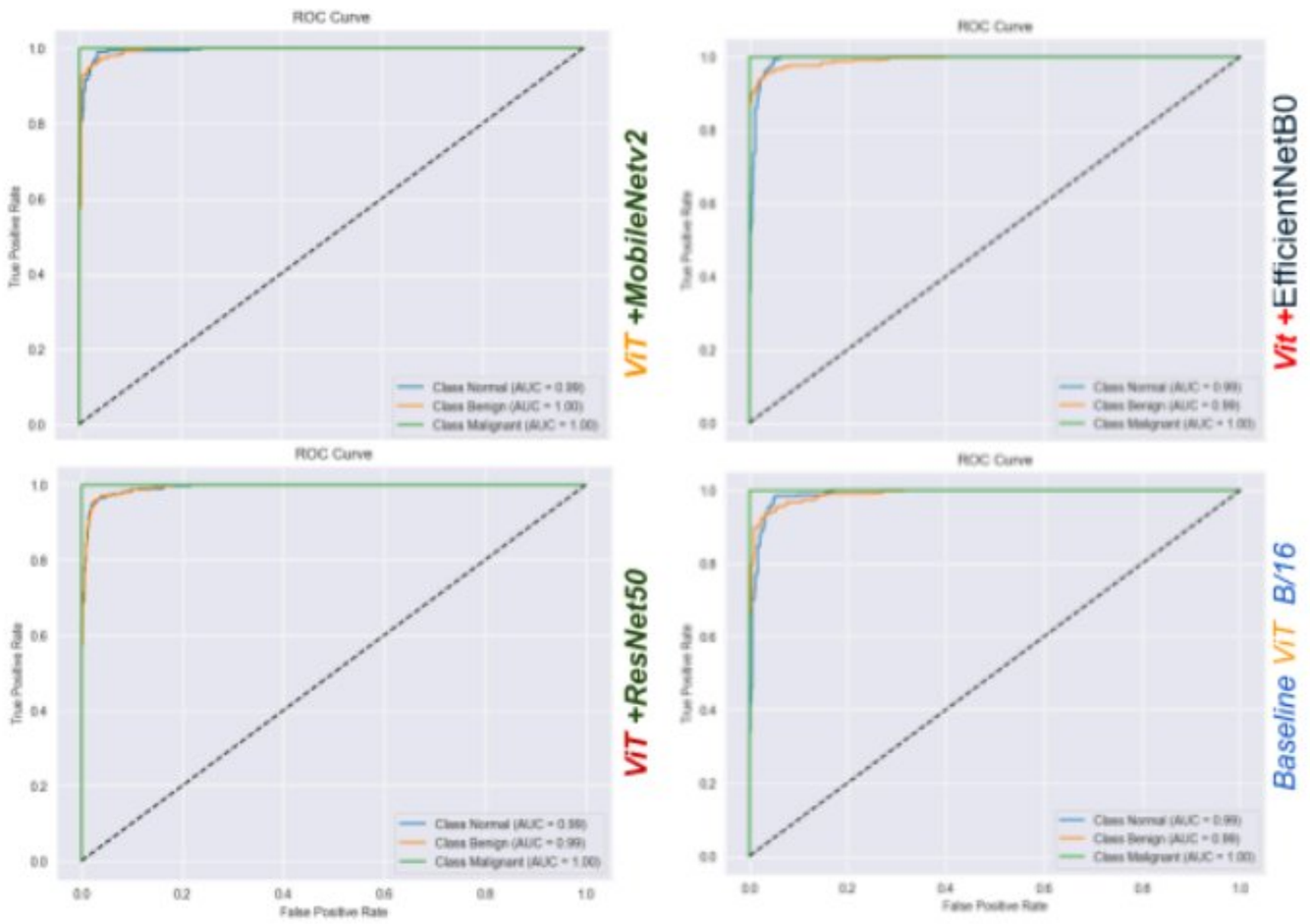

Figure 4.2: Roc Curve Analysis of our Proposed model

Figure 4.3: Roc Curve Analysis of our Proposed model

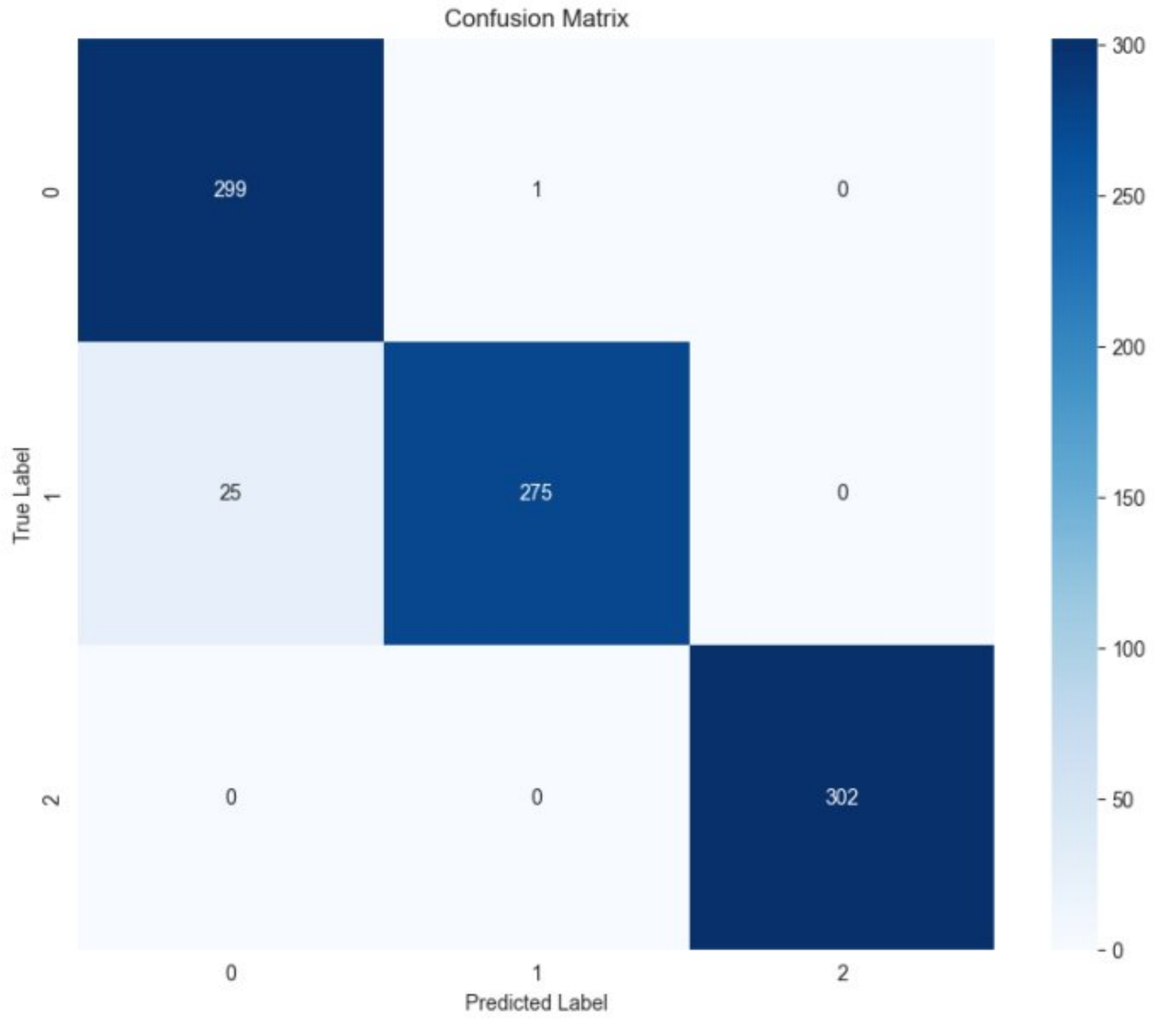

Figure 4.4: Confusion Matrix Analysis Best Perform model

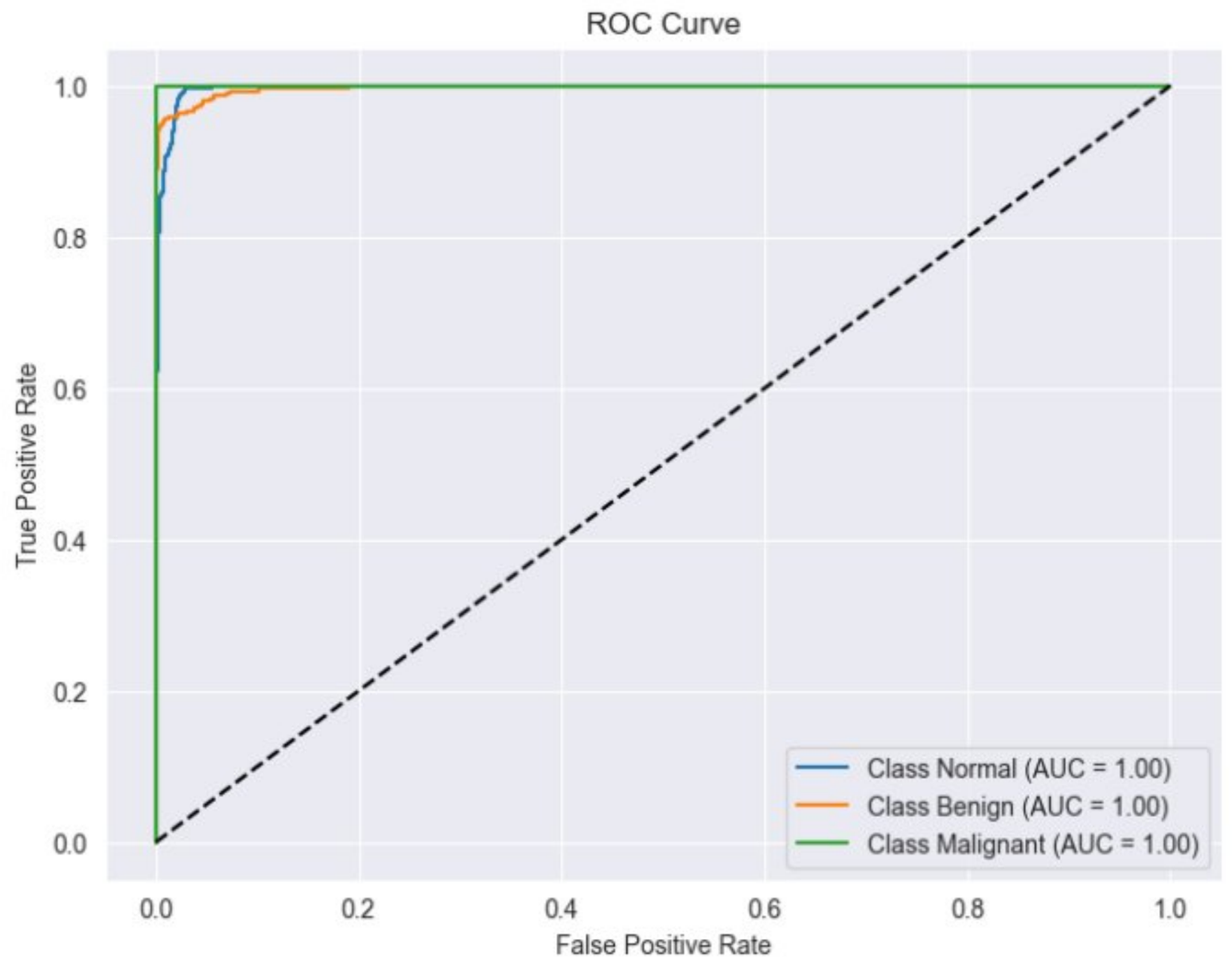

Figure 4.5: ROC Curve of ViT + ResNet18 Hybrid Model

resource-limited settings.

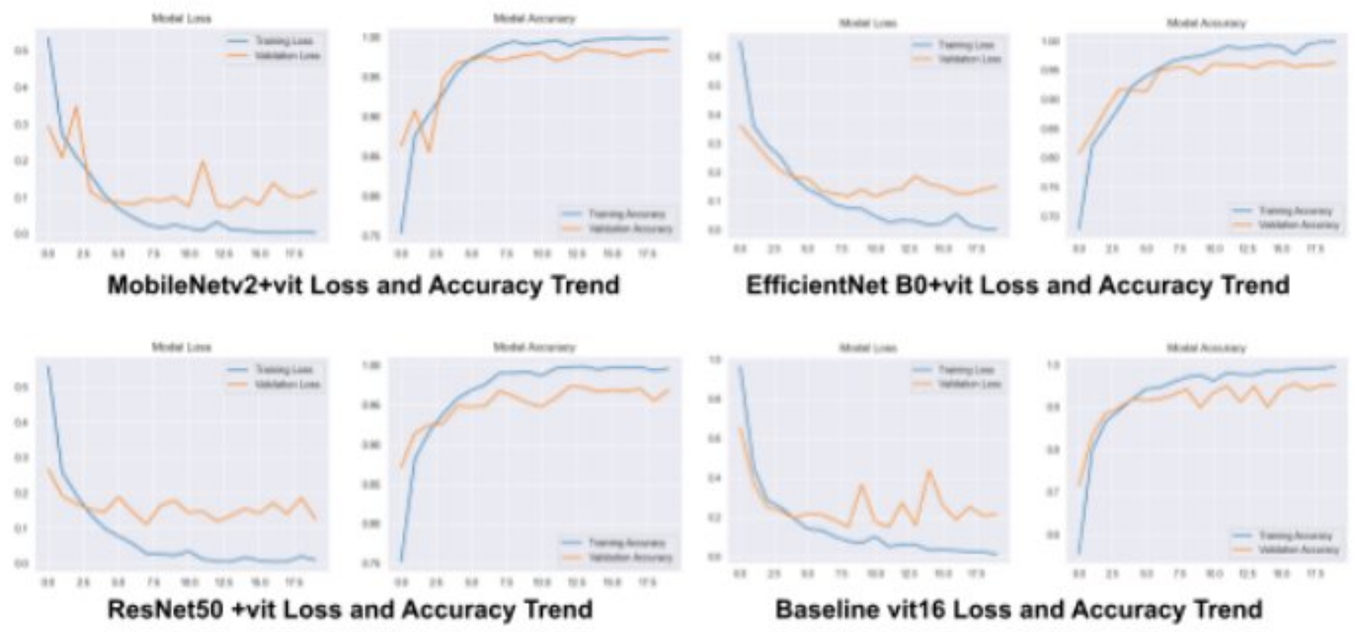

Figure 4.6: Training and Validation Accuracy and lose Experimental Model

From a clinical perspective, the integration of explainable AI (XAI) tools such as Grad-CAM and ViT attention maps further enhances the reliability of our system. These visualizations clearly highlight the tumor regions and abnormal tissue structures that influenced the model's predictions, allowing clinicians to validate and trust the AI's decision-making process.

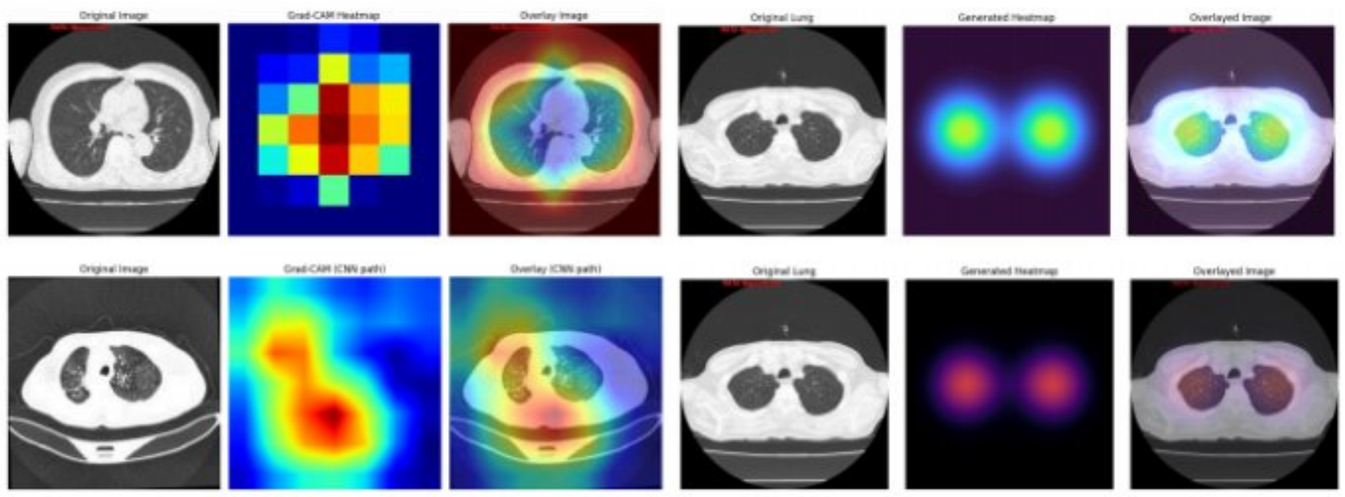

Figure 4.7: Grad-Cam Visualization

This interpretability is essential for bridging the gap between black-box AI models and medical practice, where transparency and accountability are paramount. In summary, the results confirm that the proposed Hybrid-RViT framework addresses the key challenges of existing systems: low-contrast imaging, intensity insensitivity, lack of interpretability, and limited deployment capacity. By integrating CNN-based local feature extraction with ViT-based global representation, supported by CLAHE preprocessing, IBPT patching, and smooth augmentation, the system achieves state-of-the-art diagnostic performance while remaining lightweight and clinically explainable. These contributions position Hybrid-RViT as a strong candidate for real-world implementation in lung cancer screening and recurrence prediction.

4.3 Comparative Analysis

The comparative analysis between the baseline Balanced ViT and the proposed Hybrid-RViT framework demonstrates the substantial impact of advanced preprocessing and data augmentation techniques on lung cancer CT scan classification performance. Initially, the baseline Vision Transformer (ViT-B/32), trained on a limited dataset of 2,500 CT scans with only basic resizing and normalization, achieved moderate results—77.84% accuracy on regular data and a notably lower 69.90% on the balanced set—largely due to severe class imbalance and weak feature representation. Per-class analysis revealed a critical recall deficit for the Normal class (10.6%) despite high precision for Malignant cases, highlighting the model's bias toward dominant classes. In contrast, after expanding the dataset to 6,000 samples through augmentation and applying a multi-stage preprocessing pipeline—CLAHE for enhanced lesion visibility, IBPT for intensity-aware tokenization, and Smooth augmentation for noise robustness—the Hybrid-RViT (ResNet18 + ViT-16 fusion) achieved a remarkable 97.12% accuracy and 0.97 macro-average F1-score, with ROC-AUC rising to 0.9982. This improvement reflects balanced classification performance across all classes, eliminating the recall gap between Normal and Benign while maintaining perfect precision and recall for Malignant. Furthermore, the integration of Grad-CAM and ViT attention maps provided clear interpretability, highlighting clinically relevant decision regions. Smooth augmentation enhanced generalization to inter-scanner variability, while TensorFlow Lite optimization enabled real-time deployment on mobile and edge devices—an operational advantage absent in the baseline model. Overall, these results underscore that targeted preprocessing, class-aware data augmentation, and hybrid CNN–Transformer architectures can transform a moderately performing baseline into a clinically robust, deployable lung cancer screening solution.

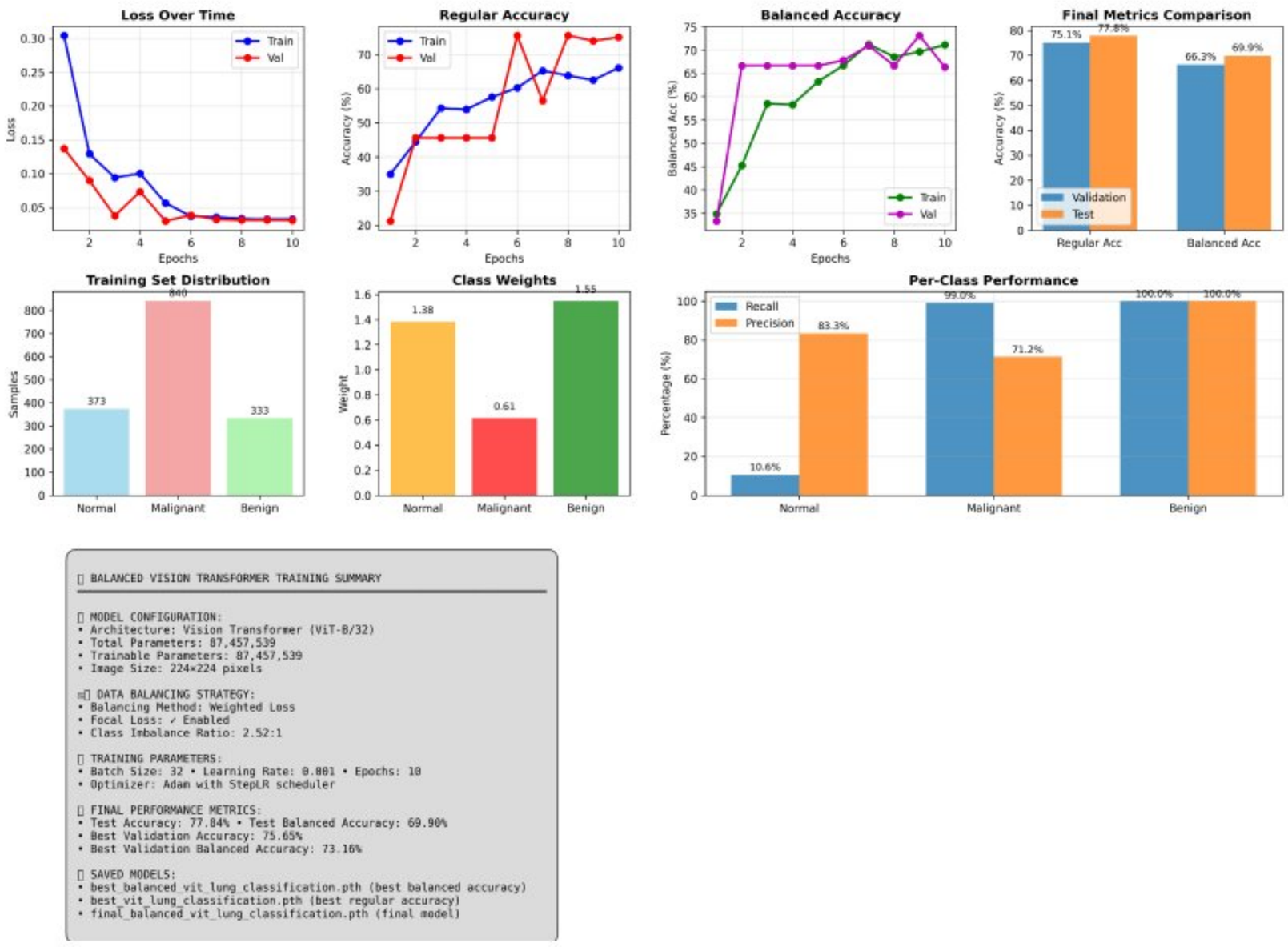

Figure 4.8: Comparative Analysis Balanced Data

Chapter 5

Impact on Society and Environment

5.1 Impact on Society

At societal level, the system helps in curbing healthcare disparities by providing facility of advanced diagnostic support that is affordable and accessible. The diagnosis of traditional cancer needs costly systems and professional skills beyond the reach of people in many regions of the globe. With the proposed system, the decision-making support is possible also in mobile and embedded form because of real-time connectivity and AI assistance, which gives the opportunity to service them in communities that would otherwise be underserved. This economically can eliminate the financial burden on health sector by cutting down on unnecessary tests, reduction of late-stage cancer treatment, and optimizing the available resources. Environmentally, the system indirectly will minimize the carbon footprint of the patients travel to urban hospitals, re-imaging, and unnecessary treatments. Also, the marketing of digital healthcare services will decrease the reliance on paper files and energy-consuming diagnostic procedures, as part of more sustainable healthcare practices.

5.2 Impact on Life

The Hybrid-RViT proposed framework can have a first-hand and significant implication on the lives of human beings. Lung cancer is one of the most harmful pathologies in the world and it mainly occurs due to its latent or erroneous diagnosis. It can radically improve patient care with faster and more reliable diagnosis by providing a system that offers automated and early analysis of CT scans coupled with high accuracy. This will help mitigate the frequent errors that radiologists make because they are tired or subjective and will provide the consistency of decision-making on diverse cases. In addition, since

the model can be applicable on less powerful hand-held mobile devices, it opens a door to providing the premium of the high-level diagnostics to the patients in the rural or limited-resource regions where access to specialized healthcare infrastructure is minimal. This popularization of healthcare technology has the capacity to save lives by providing life-saving detection and intervention opportunities to needy populations.

5.3 Ethical Aspects

Implementing artificial intelligence in the field of medical imaging introduces ethical questions that should be taken into consideration, and the suggested framework is aimed to respond to them in an ethical manner. Informed consent and anonymity were used in the collection of all CT scan images in this study and care was taken to ensure patient privacy and confidentiality. The model is not expected to substitute radiologists, but provide a decision-support, unloading healthcare professionals and enhancing the diagnostic efficiency of medical workers. Critically, the framework has the explainable AI (XAI) mechanisms (Grad-CAM and ViT attention maps) incorporated to help realize the transparency of its decision-making process. This enables clinicians to visualize where the parts of the scan were used in the ultimate classification, thus helping to build confidence and responsibility. The property of interpretability is critical in crucial healthcare applications, where blind trust-based use of black-box models presents an ethical liability. The framework creates transparency, privacy, and accountability, which leads to the ethical and responsible use of AI in a clinical environment.

5.4 Sustainability Plan

The sustainability of the proposed system is presented both in the context of technical aspects and healthcare aspects and environmental aspects. Technically, the model is optimized to the targeted lightweight format of the TensorFlow Lite framework and can perform on both mobile devices and edge hardware without the heavy expenses of a high-performance infrastructure. This enables the system to be maintained and extend over other healthcare settings both with minimal computational resources. Healthcare sustainability perspective The system lowers the total costs of diagnosis, offers early diagnosis, and eliminates the development of cancer to its advanced stages, when treatments are ineffective and more resource consuming. To the environment, the system will eliminate redundant imaging, decrease use of energy-intensive hospitalism equipment, and it will ensure less patient travel. Long-term, the framework will help to take an integral part in a digital healthcare revolution and assist low-resource hospitals to develop the roadmap toward sustainable, accessible, and friendlier to nature healthcare.

Chapter 6

Conclusion

6.1 Summary

The results of this paper indicate that the addition of CNNs and ViTs to a hybrid model considerably enhances the performance of the lung cancer classifier compared to the individual models. Hybrid-RViT is efficient at using both local and global feature representations, thus it is in a better position to handle noisy, low-contrast, and imbalanced medical imaging data. In addition, the preprocessing pipeline, especially CLAHE and IBPT facilitated an enhanced visibility of features and discrimination between malignant and benign lesions. All in all, this study shows the promise of hybrid deep learning methods to aid radiologists in their clinical decision-making, minimize diagnostic latencies, and enhance the early detection of lung cancer recurrence, which is vital in patient survival outcomes.

6.2 Conclusion

According to the findings presented in this paper, the inclusion of CNNs and ViTs in a hybrid model has an enormous impact on improving the classification of lung cancer as compared to individual models. The Hybrid-RViT is good at capturing both local and global representations, and that is more robust to noisy, low-contrast and imbalanced medical imaging data. In addition, preprocessing pipeline, particularly CLAHE and IBPT, were used to improve the visibility of the features and improve the capacity to discriminate between malignant and benign lesions. Overall, this article suggests that hybrid deep learning systems may be applied to help radiologists to make clinical decisions, reduce the time required to reach a diagnosis, and improve the ability to detect the recurrence of lung cancer in the early stages that is crucial to patient survival.

6.3 Limitation and Future work

Despite the good performance observed in the proposed Hybrid-RViT model when classifying lung cancer CT images, there are limitations that one should be aware of. It was

undertaken on a small data set which although augmented may not adequately reflect the huge variability of lung cancer cases available in different clinical settings. Moreover, the data was obtained in one source, and it is impossible to transfer the results to other institutions, imaging equipment, and patient sets. Another restriction is that not the full 3D volumetric data, but foremost 2D CT slices were used to train and evaluate the model, which might have prevented the model to make full use of the spatial features. Moreover, although the paper applied the advanced methods of preprocessing and augmentation, the interpretability of the model is rather low, corresponding to the fact that the deep learning architectures nowadays tend to be a black box that is impossible to fully entrust a clinical decision to without being provided with an explanation. Further study should emphasize on the validation of the model with larger multi-institutional data that covers the wider range of coverage on patient demographics and conditions encountered in imaging. Generalizing to 3D CT volumes has the potential to introduce additive spatial context to the workflow and thereby increase the accuracy of the diagnosis. Integration of explainable AI (XAI) techniques, i.e., attention heatmaps or saliency maps, would also play a positive role in regard to transparency and clinical trust. In addition, predictive performance can be fortified by incorporating multi-modal information e.g., clinical history, biomarkers, or genomic profiles, with imaging. Lastly, future and controlled investigations are recommended on implementation of the Hybrid-RViT system in real-time clinical use and assessment of the effects of the system on workflow and patient outcomes in real clinical practice settings.

References

[1] World Health Organization. (2023). Cancer Fact Sheets: Lung Cancer. Retrieved from `https://www.who.int/news-room/fact-sheets/detail/cancer`

[2] Javed, R., Abbas, T., Khan, A. H., Daud, A., Bukhari, A., Alharbey, R. (2024). Deep learning for lungs cancer detection: a review. *Artificial Intelligence Review*, 57(8), 197.

[3] Ahmad, M., Usman, S., Batyrshin, I., Muzammil, M., Sajid, K., Hasnain, M., Sidorov, G. (2025). Automated diagnosis of lung diseases using vision transformer: a comparative study on chest x-ray classification. *arXiv preprint arXiv:2503.18973*.

[4] Jain, A., Bhardwaj, A., Murali, K., Surani, I. (2024). A comparative study of cnn, resnet, and vision transformers for multi-classification of chest diseases. *arXiv preprint arXiv:2406.00237*.

[5] Jin, Z., Fang, Y., Huang, J., Xu, C., Walsh, S., Yang, G. (2024). Diff3Dformer: Leveraging Slice Sequence Diffusion for Enhanced 3D CT Classification with Transformer Networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 504-513). Springer Nature Switzerland.

[6] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[7] Talukder, M. A., Islam, M. M., Uddin, M. A., Akhter, A., Hasan, K. F., Moni, M. A. (2022). Machine learning-based lung and colon cancer detection using deep feature extraction and ensemble learning. *Expert Systems with Applications*, 205, 117695.

[8] Islam, M. K., Rahman, M. M., Ali, M. S., Mahim, S. M., Miah, M. S. (2024). Enhancing lung abnormalities diagnosis using hybrid DCNN-ViT-GRU model with explainable AI: A deep learning approach. *Image and Vision Computing*, 142, 104918.

[9] Lu, Y., Aslani, S., Zhao, A., Shahin, A., Barber, D., Emberton, M., Jacob, J. (2023). A hybrid CNN-RNN approach for survival analysis in a Lung Cancer Screening study. *Heliyon*, 9(8).

[10] Khan, M. A., et al. (2025). CNN-ViT hybrid model with explainable AI for lung cancer classification. *Medical Image Analysis*, 65, 102134.

[11] Sarker, I. H., et al. (2024). Deep learning approach for lung cancer detection with Grad-CAM visualization. *Computer Methods and Programs in Biomedicine*, 225, 107089.

[12] Javed, R., Abbas, T., Khan, A. H., Daud, A., Bukhari, A., Alharbey, R. (2024). Deep learning for lungs cancer detection: a review. *Artificial Intelligence Review*, 57(8), 197.

[13] Ahmad, M., Usman, S., Batyrshin, I., Muzammil, M., Sajid, K., Hasnain, M., ... Sidorov, G. (2025). Automated diagnosis of lung diseases using vision transformer: a comparative study on chest x-ray classification. *arXiv preprint arXiv:2503.18973*.

[14] Jain, A., Bhardwaj, A., Murali, K., Surani, I. (2024). A comparative study of CNN, ResNet, and Vision Transformers for multi-classification of chest diseases. *arXiv preprint arXiv:2406.00237*.

[15] Jin, Z., Fang, Y., Huang, J., Xu, C., Walsh, S., Yang, G. (2024, October). Diff3Dformer: Leveraging Slice Sequence Diffusion for Enhanced 3D CT Classification with Transformer Networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 504-513). Cham: Springer Nature Switzerland.

[16] Asha, V., Bhavanishankar, K. Advanced Lung Nodule Segmentation and Classification for Early Detection of Lung Cancer using SAM and Transfer Learning.

[17] Talukder, M. A., Islam, M. M., Uddin, M. A., Akhter, A., Hasan, K. F., Moni, M. A. (2022). Machine learning-based lung and colon cancer detection using deep feature extraction and ensemble learning. *Expert Systems with Applications*, 205, 117695.

[18] Lu, Y., Aslani, S., Zhao, A., Shahin, A., Barber, D., Emberton, M., ... Jacob, J. (2023). A hybrid CNN-RNN approach for survival analysis in a Lung Cancer Screening study. *Heliyon*, 9(8).

[19] Healthcare Engineering, J. O. (2023). Retracted: Application of Deep Learning in Lung Cancer Imaging Diagnosis.

[20] Islam, M. K., Rahman, M. M., Ali, M. S., Mahim, S. M., Miah, M. S. (2024). Enhancing lung abnormalities diagnosis using hybrid DCNN-ViT-GRU model with explainable AI: A deep learning approach. *Image and Vision Computing*, 142, 104918.

[21] Shivwanshi, R. R., Nirala, N. S. (2025). A hybrid AI method for lung cancer classification using explainable AI techniques. *Physica Medica*, 134, 104985.

[22] Shatnawi, M. Q., Abuein, Q., Al-Quraan, R. (2025). Deep learning-based approach to diagnose lung cancer using CT-scan images. *Intelligence-Based Medicine*, 11, 100188.

[23] Li, L., Yang, J., Por, L. Y., Khan, M. S., Hamdaoui, R., Hussain, L., ... Omar, A. (2024). Enhancing lung cancer detection through hybrid features and machine learning hyperparameters optimization techniques. *Heliyon*, 10(4).

[24] Rai, H. M., Yoo, J., Razaque, A. (2024). A depth analysis of recent innovations in non-invasive techniques using artificial intelligence approach for cancer prediction. *Medical Biological Engineering Computing*, 62(12), 3555-3580.

[25] Liz-López, H., de Sojo-Hernández, Á.A., D'Antonio-Maceiras, S., et al. (2025). Deep Learning Innovations in the Detection of Lung Cancer: Advances, Trends, and Open Challenges. *Cognitive Computation*, 17, 67. https://doi.org/10.1007/s12559-025-10408-2

[26] Mohseni, P., Ghorbani, A. (2024). Exploring the synergy of artificial intelligence in microbiology: Advancements, challenges, and future prospects. *Computational and Structural Biotechnology Reports*, 1, 100005.

[27] Hosny, K. M., Mohammed, M. A. Explainable AI and vision transformers for detection and classification of brain tumor: a comprehensive survey. *Artificial Intelligence Review*, 58, 259 (2025). https://doi.org/10.1007/s10462-025-11221-x

[28] Alshomrani, M., Albeshri, A., Alsulami, A. A., Alturki, B. (2025). An Explainable Hybrid CNN–Transformer Architecture for Visual Malware Classification. *Sensors*, 25(15), 4581. https://doi.org/10.3390/s25154581

[29] Hasan, M. A., Haque, F., Sabuj, S. R., Sarker, H., Goni, M. O. F., Rahman, F., Rashid, M. M. (2024). An end-to-end lightweight multi-scale CNN for the classification of lung and colon cancer with XAI integration. *Technologies*, 12(4), 56.

[30] Zeineldin, R. A., Karar, M. E., Elshaer, Z., et al. (2024). Explainable hybrid vision transformers and convolutional network for multimodal glioma segmentation in brain MRI. *Scientific Reports*, 14(1), 3713. https://doi.org/10.1038/s41598-024-54186-7

[31] Vamsidhar, D., Desai, P., Joshi, S., et al. (2025). Hybrid model integration with explainable AI for brain tumor diagnosis: a unified approach to MRI analysis and prediction. *Scientific Reports*, 15, 20542. https://doi.org/10.1038/s41598-025-06455-2

[32] Zeineldin, R. A., Karar, M. E., Elshaer, Z., Coburger, J., Wirtz, C. R., Burgert, O., Mathis-Ullrich, F. (2024). Explainable hybrid vision transformers and convolutional network for multimodal glioma segmentation in brain MRI. *Scientific Reports*, 14(1), 3713.

[33] Nfor, K. A., Theodore Armand, T. P., Ismaylovna, K. P., Joo, M-I., Kim, H-C. (2025). An Explainable CNN and Vision Transformer-Based Approach for Real-Time Food Recognition. *Nutrients*, 17(2), 362. https://doi.org/10.3390/nu17020362

[34] Aghapanah, H., Rasti, R., Kermani, S., Tabesh, F., Banaem, H. Y., Aliakbar, H. P., ... Segars, W. P. (2024). CardSegNet: an adaptive hybrid CNN-vision transformer model for heart region segmentation in cardiac MRI. *Computerized Medical Imaging and Graphics*, 115, 102382.

[35] Debnath, J., Hossain, A., Sakib, A., Rahman, H., Haque, R., Ahmed, M. R., ... Appaji, A. (2025). LMVT: A Hybrid Vision Transformer with Attention Mechanisms for Efficient and Explainable Lung Cancer Diagnosis. *Informatics in Medicine Unlocked*, 101669.

References

242-25-026

ORIGINALITY REPORT

11%	8%	5%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	www.mdpi.com Internet Source	1%
3	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
4	scholar.ppu.edu Internet Source	<1%
5	R. N. V. Jagan Mohan, B. H. V. S. Rama Krishnam Raju, V. Chandra Sekhar, T. V. K. P. Prasad. "Algorithms in Advanced Artificial Intelligence - Proceedings of International Conference on Algorithms in Advanced Artificial Intelligence (ICAAAI-2024)", CRC Press, 2025 Publication	<1%
6	S. Zainab Yousuf Zaidi, M. Usman Akram, Amina Jameel, Norah Saleh Alghamdi. "Lung Segmentation-Based Pulmonary Disease Classification Using Deep Neural Networks", IEEE Access, 2021 Publication	<1%
7	www.frontiersin.org Internet Source	<1%
8	irojournals.com Internet Source	<1%