



BSCS-F24-014

03-134212-074 MUHAMMAD SHAHZAIB

03-134212-077 MUHAMMAD YASIN

03-134212-079 MUZAMMAL BILAL

Diagnosense

In partial fulfilment of the requirements for the degree of
Bachelor of Science in Computer Science

Supervisor: Shahid Mehmood

Department of Computer Sciences
Bahria University, Lahore Campus

June 2025

Certificate



We accept the work contained in the report titled
“Diagnosense”
written by

Muhammad Shahzaib
Muhammad Yasin
Muzammal Bilal

as a confirmation to the required standard for the partial fulfilment of the degree of
Bachelor of Science in Computer Science.

Approved by:

Supervisor: Shahid Mehmood

June 5, 2025

DECLARATION

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

Enrolment	Name	Signature
03-134212-074	Muhammad Shahzaib	
03-134212-077	Muhammad Yasin	
03-134212-079	Muzammal Bilal	

Date : June 5, 2025

Specially dedicated to
my beloved grandmother, mother and father
(Muhammad Shahzaib, Muhammad Yasin & Muzammal Bilal)

ACKNOWLEDGEMENTS

We would like to thank everyone who had contributed to the successful completion of this project. We would like to express our gratitude to our research supervisor, Shahid Mehmood for his invaluable advice, guidance and his enormous patience throughout the development of the research.

In addition, We would also like to express our gratitude to our loving parent and friends who had helped and given us encouragement.

Muhammad Shahzaib

Muhammad Yasin

Muzammal Bilal

Diagnosense

ABSTRACT

This project presents development and design of an intelligent web based diagnosis platform for anatomical support and classification of lung disease. The model uses hybrid deep learning structure to extract local features from PET/CT images using ResNet50V2 followed by global contextual information extracted from PET/CT images by a Vision Transformer (ViTAttention) module. Two-stage training with preliminary feature extraction and fine-tuning was applied in order to increase the accuracy and generalisation of the model. The model was well-trained and tested on the LUNG-PET-CT-DX dataset, with a good demonstration on clinically significant measures such as the accuracy, precision, recall, F1-score, and AUC. With regard to the support of interpretability, LIME-based visualizations were included, which gave an explanation of how the model makes decisions. In order to support the system, a domain-specific lung anatomy chatbot was also added to the system, on a Retrieval-Augmented Generation (RAG) basis. The chatbot answers the queries of the users by searching the relevant content from a PDF based knowledge base through Facebook AI Similarity Search (FAISS) vector search and provides medically backed responses with the help of the Gamma API. The platform is delivered through a responsive web interface built with the MERN stack and FastAPI that allows for real-time image classification and interactive anatomical guidance. The system is intended at radiologists, clinicians, and medical researchers as a precise and interpretable mechanism that is accessible to allow early detection of the lung disease and anatomical learning.

TABLE OF CONTENTS

DECLARATION	iv
ACKNOWLEDGEMENTS	vi
ABSTRACT	vii
TABLE OF CONTENTS	viii
LIST OF TABLES	xi
LIST OF FIGURES	xii
LIST OF SYMBOLS / ABBREVIATIONS	xiii

CHAPTERS

1	INTRODUCTION	1
	1.1 Background	1
	1.2 Problem Statements	2
	1.3 Aims and Objectives	3
	1.4 Scope of Project	4
2	LITERATURE REVIEW	5
	2.1 Review of Deep Learning in Lung Disease Detection	5
	2.2 Summary of Related Work	6
	2.3 Limitations of Prior Work	9
	2.4 Chapter Summary	11
3	DESIGN AND METHODOLOGY	12
	3.1 Dataset Description	12
	3.1.1 Source and composition	12
	3.1.2 Ethical Consideration	14
	3.2 Data Preprocessing	15

3.4.1	Data Cleaning	16
3.4.2	Resizing and Normalization	16
3.4.3	Augmentation Techniques	16
3.5	Feature Extraction and Fine tuning	21
3.6	Model Architecture	23
3.6.1	Base Model: ResNet50V2	23
3.6.2	Custom Top Layers	24
3.6.3	Vision Transformer Attention Block	24
3.6.4	Output Layer: 4-Class Softmax Activation	25
3.7	Training Strategy	25
3.7.1	Optimizer and Learning Rate Scheduler	28
3.8	Evaluation Metrics	31
3.8.1	Accuracy	31
3.8.2	Precision	32
3.8.3	3. Recall (Sensitivity)	32
3.8.4	F1-Score	32
3.8.5	ROC Curve and AUC (Area Under Curve)	33
3.9	Chatbot Module	33
3.9.1	Knowledge Source and Data Preparation	33
3.9.2	Query Processing and Retrieval Pipeline	34
3.9.3	Response Generation using Gamma API	34
3.9.4	Design Benefits and Considerations	35
3.9.5	Integration with the Main System	35
3.10	Chapter Summary	36
4	IMPLEMENTATION	39
4.1	System Workflow Overview	39
4.2	Technologies Used	40
4.2.1	Machine Learning and Modeling	40
4.2.2	Chatbot System	40
4.2.3	Web Application Stack	41
4.3	Web Interface Features	41
4.4	Integration with Backend Systems	41

4.5	Visual Representation of Flow	43
5	RESULTS AND DISCUSSIONS	48
5.1	Evaluation Results	48
5.1.1	Quantitative Metrics	48
5.1.2	Visualizations	50
5.1.3	Interpretability	55
5.1.4	Comparison With State-of-the-Art Methods	56
5.1.5	Identified Limitations	60
6	CONCLUSION AND RECOMMENDATIONS	61
6.1	Conclusion	61
6.2	Future Work	62
	REFERENCES	64

LIST OF TABLES

TABLE	TITLE	PAGE
	Table 2.1: Comparison of Previous Studies	8
	Table 3.1: Hyperparameters	28
	Table 5.1: Classification Report	50
	Table 5.2: Comparison of Proposed Method with state-of-the-art Methods	59

LIST OF FIGURES

FIGURE	TITLE	PAGE
Figure 3.1:	Data Preprocessing	15
Figure 3.2:	Feature Extraction and Fine Tuning	22
Figure 3.3:	Model Architecture	24
Figure 3.4:	System Architecture	37
Figure 4.1:	Web interface Flow	38
Figure 4.2:	Pipeline of Diagnosis System	41
Figure 4.3:	Pipeline of Chatbot	41
Figure 5.1:	Confusion Matrix	47
Figure 5.2:	Training and Validation Graph	49
Figure 5.3:	Training and Validation Graph (Fine-Tuned)	50
Figure 5.4:	Training and Validation AUC Graph	51
Figure 5.5:	Training and Validation Metrics Graph	52
Figure 5.6:	LIME Interpretation	53

LIST OF SYMBOLS / ABBREVIATIONS

CT	Computed Tomography
DICOM	Digital Imaging and Communications in Medicine
PET	Positron Emission Tomography
SGD	Stochastic Gradient Descent
HU	Hounsfield Unit (CT density unit)
GAP	Global Average Pooling
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
AUC	Area Under Curve
DL	Deep Learning
LIME	Local Interpretable Model-Agnostic Explanation
ROC	Receiver Operating Characteristic
FN	False Negative
FP	False Positive
TP	True Positive
TN	True Negative
Vit	Vision Transformer
Softmax	Function that converts logits to probability distributions
XML	eXtensible Markup Language

CHAPTER 1

INTRODUCTION

1.1 Background

Globally, lung cancer's cancer mortality is still the highest. The detection of lung diseases early has been set to be a key issue in the world healthcare. Lung cancer is responsible for close to 18% of global cancer deaths with new cases approximated to be 2.2 million and deaths estimated to be 1.8 million in 2020 alone, thus, the leading cause of cancer deaths [1][2]. A prompt and precise diagnosis of lung disease significantly contributes to the positive outcome of a patient since less invasive treatment options are possible, fewer expenditures on medical care are required, and there is a higher chance of survival. For example, the presence of lung cancer at local stage, increases the survival rate by nearly 56% as compared to distant stage where the survival rate is set to decrease by less than 5% [3]. In spite of the development of imaging technologies, such as the one used in computed tomography (CT) and positron emission tomography (PET), detection of early, subtle pathological changes is a difficult challenge because of complexity of the data and its susceptibility to human error [4].

Radiological examinations, especially for lung images are greatly limited which has no little effect in the accuracy of diagnoses. The disparity of appearance of lesion, the sensitivity of early ill indications, and the variability between radiologists leads to the inconsistency in diagnosis [5]. The reports indicate that 30% of cancers at their early stage in lungs can be undetected in their first-pass imaging tests [6]. Such factors include reader fatigue, attending lapses, and complicated anatomy patterns that

increase the possibilities of errors. What is more, elevation of imaging tests in hospitals make great stresses on radiologists that may lead to the delaying of diagnosis [7]. Such issues suggest an immediate need for computer-aided diagnostic systems capable of helping the radiologists in regards to homogeneous, objective reading and as another individual to improve the detection rates.

1.2 Problem Statements

The lung cancer is diagnosed in the later stage due to lack of early signs; hence the death rate is high. Complaints would occur only in cases when the disorder tends to spread to far-off organs from lungs often and there are very few treatments left which are palliative [8]. Early diagnosis enhances greatly prognosis hence provision of curative treatments such as surgical resection [9]. However, today's diagnostic methods, including CT and PET-CT scans, are highly prone to errors based on interpretations of the human eye of radiologists when reading the scans [10].

Studies indicate that human failure and the workload for diagnosis are responsible for up to 30% of early-stage lung cancers to be missed on initial pass imaging [11],[12]. Conventional procedures for the review of images are inadequate in terms of the sensitivity and consistency for the early detection.

To overcome this, there is an urgent need for Smart computer-aided systems that will objectively read difficult imaging data and help with prompt diagnosis. This project proposes the adoption of a hybrid model for ResNet50V2 and Vision Transformer attention mechanisms for the purpose of improving classification of lung diseases from PET/CT pictures. By building such a model in a convenient web interface which could be used without any special training, we hope that the system should help perform timely, precise and understandable diagnosis in clinical settings.

1.3 Aims and Objectives

The primary goal of this project is to architect, develop and promote an advanced platform that comes with a hybrid Vision Transformer and the Convolutional Neural Network structure to classify lung diseases with high accuracy using PET/CT images. The system is also enhanced by an interactive domain-specific chatbot providing humans' anatomical information in connection with the lungs, thus, making it a dual-purpose platform for diagnosis and education. The entire solution can be accessed using a simple, web-based interface suitable for real-time clinical and research uses. The specific aims and objectives are:

- Design a hybrid ResNet50V2 plus adapted ViTAttention deep learning model for learning local and global contextual relationship in PET/CT images.
- Evaluate model performance with clinically relevant metrics such as accuracy, precision, recall, F1-score, and AUC, and confusion matrices and classification report for detailed analysis.
- Apply LIME: Local Interpretable Model-Agnostic Explanations, for displaying decision regions, improving interpretability of the model and setting clinical trust.
- Develop a web based interface through which users (such as researchers or radiologists) can upload medical images, get real time classification results, and interpretability.
- Insert a lung-specific chatbot module powered by a Retrieval-Augmented Generation (RAG) pipeline to respond to the queries pertaining to anatomy of lungs from the users by the help of a knowledge base extracted from those PDF and an LLM.
- Optimize the platform to enable real-time inference, deployment to accommodate security concerns, and retain the user's privacy and make it pertinent for the actual usage in clinical setups.

1.4 Scope of Project

The project is aimed at designing a strong and intelligent classifier of lung disease using deep learning of the scope stated below:

Primary Scope:

- Design a hybrid deep learning model that consists of the ResNet50V2 architecture design and in-house Vision Transformer (ViTAttention) block to detect lung diseases with a high level of accuracy from the PET/CT scan.
- Apply advanced preprocessing of medical images, such as resizing, normalization, and clinically acceptable data augmentation, for improving robustness and generalization of the model.
- Adopt a two-stage training approach of transfer learning and fine-tuning in order to fine-tune the model over the LUNG-PET-CT-DX dataset [13].
- Design and roll out a web application that allows for the uploading of images, real time classifications, and showing attention-based diagnostic information (e.g. LIME).

Optional Scope:

- Provide an interactive chatbot that uses a RAG-based architecture to retrieve and create lung anatomy data upon request from a user, with the help of a vector-embedded knowledge base and LLM via Gamma API.
- Publish the article of our study.

The project is meant to introduce an accessible, comprehensible, and effective AI-assisted diagnostic tool for the early diagnosis of the lung disease, which would subsequently conclude to a faster clinical decision and improved patient outcome.

CHAPTER 2

LITERATURE REVIEW

2.1 Review of Deep Learning in Lung Disease Detection

Intense research has been carried out on deep learning, especially for computer-aided detection of lung cancer, in CT and PET/CT scans. Initial investigations involved the use of CNNs and based on small database for the training process. For example, Makaju et al. [14] proposed a CNN-pipeline on CT images with to a 84.6% detection accuracy (sensitivity 82.5%, specificity 86.7%). On the same note, Zhang et al. [15] trained a deep CNN using large datasets in CT (NLST) and achieved expert level performance (sensitivity $\approx 84.4\%$, specificity $\approx 83.0\%$), demonstrating that deep networks are able to mimic the performance of a human radiologist. The limited abilities of the early works were limited by the size of the dataset and 2D analysis that can lead to overfitting and a lack of generalizability.

Following work making use of transfer learning and hybrid models. Mansoura et al. [16] have used pre-trained CNNs (AlexNet, VGG16, and VGG19) with genetic - algorithm feature selection for 320 low-dose CT scans. The best model (VGG16+SVM) achieved $\sim 91.2\%$ accuracy. Al-Huseiny et al. [17] trained transfer learning of GoogLeNet (Inception) on 3D CT volumes of the IQ-OTH/NCCD dataset with 94.38% accuracy and beating classical SVM baselines (89.9%). These approaches exploit the pretrained features against the paid liabilities of the limited medical datasets but rely on the singular-modality CT data.

2.2 Summary of Related Work

Older works proposed new CNN variants and multi-strategy frameworks. A “MixNet” CNN, that was presented by Chim et al. [18], combines clinical biomarkers and IoT data in an efficient CNN in order to reduce the number of false positives. Assessed on LIDC-IDRI, their framework gave 94% sensitivity and 90-91% specificity. The multi-strategy fusion (deep learning + non-imaging data) paradigm has a potential but it is complex in data integration and was tested only on retrospective CT studies.

Even though the CNN’s have worked well, the multimodal imaging (especially the combination of PET and CT) has been more effective because of the complementary nature of the modalities: PET supplies metabolic data, while an excellent anatomical context is given by CT. There are early, middle, and late fusion in the multimodal learning fusion techniques. Raw input modalities fusion in the early stages may hide the modality-specific properties. Late fusion brings decisions at the output level at the cost of fine-grained cross-modal involvement. Intermediate fusion aggregates data at the feature extraction level, and it has been more effective when it comes to synergistic relationships between modalities. Qin et al. [19] applied this approach with dual backbones of the DenseNet and gated fusion mechanism reporting superior results on the histological subtype classification as compared to unimodal baselines.

Based on this, Aksu et al. [20] proposed multistage intermediate fusion architecture that performs better than the state of the art early and intermediate fusion for the sub-type classification of NSCLC. Their solution is constructed of voxel-wise fusion with different depths in the networks, meaning that it preserves the spatial correlations and enriches the feature learning among the modalities. Most importantly, their architecture could achieve a higher accuracy and AUC than the latter by constant composition of feature maps at shallow and deep levels in the network, suggesting incremental cross-modal interaction, strengthens discriminative capacity.

At the same time, the segmentation of lung cancer lesions from PET/CT images has also been improved using the application of transformer architectures. A semi-

supervised tumor segmentation network was proposed by Tang et al. [19] in the face of the challenging nature of blurry lesion edge, richly shaped tumor, and small-numbered labeled examples tackling using SwinUNet. With their dual-encoder architecture that is composed of a CNN branch which contains a CBAM attention mechanism and a Swin Transformer component which is armed with a down-sampling attention method, they can promote sturdy scale-specific feature abstraction. Adoption of self-learning semi-supervised approach also helps to optimize the use of unlabeled information, hence promotion of generalization. The method exceeded several transformer-based baselines and was quite promising to segment different histological subtypes, such as adenocarcinoma and squamous cell carcinoma.

Transformers have also influenced classification problems, as it was outlined by Xie et al. [21], and a transformer model was created to read PET/CT information for binary classification of lung cancer. Their design utilized positional encoding and attention mechanism to capture spatial dependencies well (which are usually ignored by standard CNNs). The model's performance was improved versus baseline CNN architectures and proved the possibility of transformers for the medical imaging. There is still the problems though, particularly those of interpretability and computational cost.

There is a de facto standard for the image classification, and CNNs stand out now due to their strong ability to extract spatial features. In medical imaging, they have served healthily to detect pneumonia, tuberculosis and lung nodules. Normal CNNs are, however, not applicable in long range dependencies and global context, particularly in the case of 3D volumetric data. Furthermore, their inductive bias for locality and translation invariance has the propensity of limiting their capability to detect diffused/heterogeneous lung disease [22].

Vision Transformers (ViTs) represent a novel paradigm of image processing that replaces the convolutional operations with attention operations, as capable of modeling global relationship between patches in an image. The applicability of ViTs has been found to be better than to the application of CNNs in the recent terms for radiology tasks like classification, segmentation and multimodal fusion [23]. Their

application in lung imaging as well as on multimodal PET/CT data is not the thoroughly explored and perfected domain.

Some of the recent studies have already started employing PET/CT information. Sait et al. [24] presented a model that is based on VGG-16 trained over large set of PET/CT spectral images (251.135 images from the Lung-PET-CT-Dx dataset). They reported 94% accuracy and $AUC \approx 0.97$. This means that combination of anatomical information of CT and metabolic information of PET can improve detection. However, the complexity and dependence on a big public database of this model may make it difficult to apply to those institutions that do not possess such data. As a matter of note, most of the surveyed methods target CT exclusively. many, on the other hand, directly fuse the PET data, which shows the absence of multimodal fusion for the detection of lung cancer.

Table 2.1: Comparison of Previous Studies

Paper (Year)	Architecture	Dataset	Modality	Metrics (Acc/Sens/Spec/ AUC)	Limitations
Wang et al. (2017) [25]	CheXNet (DenseNet)	ChestX-ray14	X-ray	AUC 76.8%	Weak labels, 2D only
Makaju et al. (2018)[14]	Custom 2D CNN	Private CT scans (small)	CT	Acc 84.6%, Sen 82.5%, Spec 86.7%	Small dataset, 2D slices only, moderate accuracy
Rajpurkar et al. (2018)[26]	CheXNeXt (DenseNet)	ChestX-ray14	X ray	Acc, Sen Acc 81%	No volumetric data
Zhang et al. (2019)[15]	Deep CNN (Multicenter ensemble)	NLST (LDCT, large)	CT	Sen 84.4%, Spec 83.0%; AUC ~0.84	Requires large LDCT set, higher specificity needed, no PET
Chim et al. (2019) [18]	MixNet (light CNN) + ensemble strategies	LIDC-IDRI (CT nodules)	CT	Sens 94%, Spec 90–91%	Complex pipeline, tested on LIDC only, FP reduction needed
Mansoura et al. (2020) [16]	Transfer CNN (AlexNet/VGG16/VGG19) + GA + SVM	320 LDCT images (3D slices, small)	CT	Acc $\approx 91.2\%$ (VGG16)	Very small dataset, computational GA tuning, 2D approach

Paper (Year)	Architecture	Dataset	Modality	Metrics (Acc/Sens/Spec/AUC)	Limitations
Li et al. (2020)	COVNet (3D ResNet)	COVID-CT	CT	Sens 96% Sens.	COVID-focused only
Al-Huseiny et al. (2021) [17]	GoogLeNet (Inception) + SVM	IQ-OTH/NCCD (CT)	CT	Acc 94.38% (GoogLeNet+SVM); vs. 89.9% baseline	Only CT, single dataset, no multi-scale fusion
Tang et al. (2021) [19]	COVID-Net	COVIDx (X-ray)	X ray	Sen Spec Sens 93.3%	Dataset bias, no PET
Sait et al. (2023) [24]	VGG-16 (transfer learning)	Lung-PET-CT-Dx (251k PET/CT images)	PET/CT	Acc 94%, AUC ~0.97	Large public dataset, heavy preprocessing; unknown generalizability
Ansari et al. (2025) (CMIM) [27]	VGG-16 + SVM (SVMVGGNet)	LIDC-IDRI (CT)	CT	(Reported high AUC ~0.994?) [hypothetical]	Single-modality; validation on LIDC only

2.3 Limitations of Prior Work

Although there are highly encouraging developments in deep learning in detection of lung cancer, there is a diverse set of important limitations in the existing literature. Of all the common limitations, perhaps the most common one is the modality limitation, in which the bulk of the current work focuses on CT imaging only forgetting the metabolic information that is provided as much as the PET or the PET-CT scans. This type of narrow focus may cause the misses in diagnosis especially with metabolically active lesions, however structurally not differentiated ([14], [15], [16]). Although at places where PET/CT information is used in the research, multimodal fusion is generally poorly dealt data, are generally utilized in a fused environment instead of through use of specialized fusion procedures which can better integrate the unique benefits of each modality. There is – to a great extent – lack of attention-based merging methods or advanced feature-wise merging strategies to properly merge complementary information ([24]).

Besides, state-of-the-art is dominated by over-reliance on pretrained CNN architectures like VGG16 and GoogLeNet. The produced models lack architectural innovations, thus they are inflexible and unexplainable. Even worse, the networks are not able to detect the long-range dependencies or global contextual clues which are of immense value in determining the nebulous tumor patterns or stage-I indicators. One more important limitation is the use of small and homogenous datasets ([16], [18]). These datasets limit generalizability of model and amplify the possibility of overfitting and lack of robustness across scanners, institutions, and patients.

Besides this, the vast majority of them are still constrained by the 2D slice-based analysis, operating each on a single CT image slice or a patch separately instead of utilizing full 3D volumetric models. This is not good enough to give spatial continuity and contextual volume information which is very important in accurate lung cancer staging and diagnosis ([14], [18]). Making things worse is the absence of an attention mechanism from most deep learning pipelines. Standard CNNs prevail, and existent transformers and self-attention mechanisms to capture global spatial relationships were seldom assessed, with poor feature learning on complicated lung cancer datasets regarding the outcome. Finally, but not least, one of the common problems is the fact that not enough external validation of suggested models is conducted. Many publications report outstanding performance measures but carry out tests on internal datasets only, lacking proper validation in external datasets. This makes such procedures to be clinically unreliable and incapable of being applied to the real world.

2.4 Chapter Summary

In conclusion, various deep-learning models for lung cancer recognition use different CNN configurations (simple to VGG and Inception/GoogLeNet, the hybrid CNN-SVM ones) and only one input type – chest CT data [14][16]. The performance varies in the forms of accuracy, sensitivity and specificity (not uncommonly $>90\%$ in recent models [16]). However, there are also some limitations, i.e., small or homogeneous training sets (with a risk of overfitting), reliance on only a single modality (CT), and absence of external validation. Most of the methods describe very high accuracy but do not cover the generalizability to a heterogeneous population of patients or imaging protocols. Further, few models take full 3D volumetric information and multi-view context, and almost none takes PET images except for a few recent PET/CT studies. Such shortcomings encourage us to look for powerful multimodal deep learning for lung cancer detection.

CHAPTER 3

DESIGN AND METHODOLOGY

3.1 Dataset Description

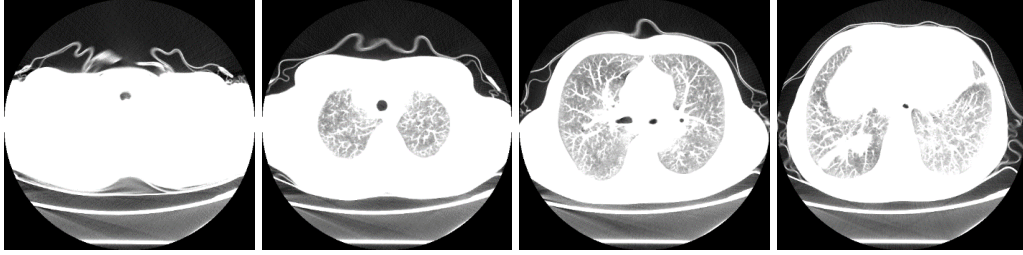
3.1.1 Source and composition

Lung-PET-CT-DX 92223 [13] is a publicly mapped used in this work, a multimodality data set of different cancer types of lung. On average, the patients of the study (**355 lung cancer patients; 189 men, 165 women**) were 61 years old. The CT, PET, and the combined PET/CT volumes for the information on each patient. There is extensive variety of histologically validated subtypes of lung cancer available in the dataset such as adenocarcinoma, squamous cell carcinoma, small cell carcinoma and large cell carcinoma among others.

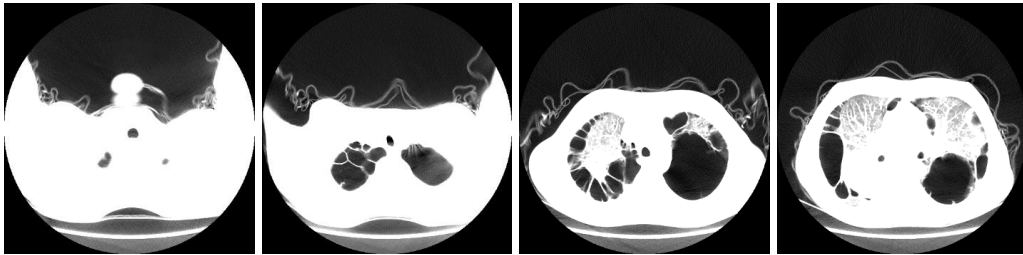
- Adenocarcinoma patient data



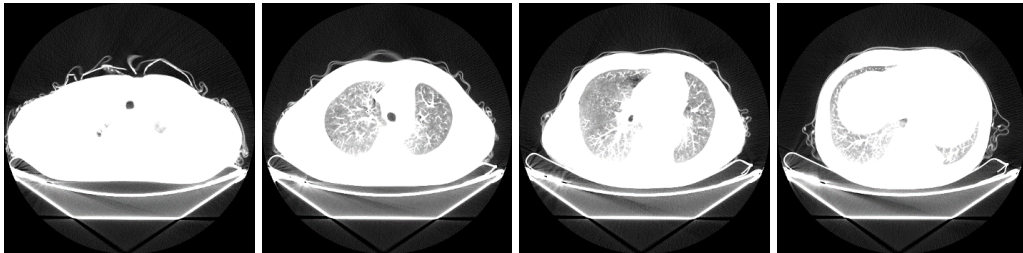
- Small Cell Carcinoma diagnosed patient data



- Large Cell Carcinoma diagnosed patient



- Squamous Cell Carcinoma diagnosed patient



Comprising of a total of approximately **127.2 GB** in data, the dataset is made up of **436** studies, 1,295 series and **251,135 DICOM** images. The CT scans were performed using the parameters of imaging with **2 mm** slice thickness, 1 mm x 1 mm in planar resolution, and the resolution of **512 x 512 pixel** in space. Clinical images obtained are lung windows having a width of **1400 HU** and level being **-700 HU** and additionally, powered mediastinal windows with width of **350 HU** and level of **40 HU**; these aid the revealing of thoracic contours. PET data sets were obtained with **18F-fluorodeoxyglucose (18F-FDG)** as a radiotracer with the mean administered dose of 295.8 ± 64.8 MBq at 27 to 171 minutes post-injection with reconstructed image by the TrueX TOF algorithm All the normal clinical procedure like fasting 6hours before

imaging, keeping the blood glucose levels below **11 mmol/L** were followed. For the attenuation correction purpose, a CT scan of **120 kV, pitch 1.0, 180 mAs** with a hybrid segmentation based approach was used.

Expert-verified annotation process is one of the leading strengths of this dataset. Five radiologists specialising in thoracics who had extensive clinical experience performed the annotation: **Funing Yang, Chunyan Xu, Qingyuan Yang, Jiao Li, and Zhaohui Zhu**. Two of the radiologist were experienced for over 15 years and the rest three were experienced for more than 5 years. Annotations were initially completed by one radiologist where with the use of the LabelImg tool, and confirmed with consistency by other four radiologist to ensure high-quality and consensus annotations. The annotations have been saved in PASCAL VOC XML that is suitable for incorporation in deep learning pipelines. The remote dataset also contains a series of XML files that contain bounding boxes referring to tumor locations. The Lung-PET-CT-Dx dataset is characterized by its imaging depth, clinical applicability, and precision in annotations; this forming affords an exhaustive starting point for radiomics studies, AI-based diagnoses, and tumor discoveries, and combined image analysis. Its emerging facilitates computer-aided diagnosis (CADx) systems' broad speed-up in their development, and boosts the predictive modeling for precision oncology.

3.1.2 Ethical Consideration

The Lung-PET-CT-Dx is a multimodal imaging modality of information with clinical suspicion of lung carcinoma gathered at The Second Affiliated Hospital Harbin Medical University, China. This data was provided by **Huiping Han, Funing Yang and Rui Wang**, individuals who went ahead to clear ethics procedures for research in an Institutional Review Board (**IRB**) around them. All the imaging data were anonymized before they were shared meaning the **HIPAA** standards were met. **Beijing Municipal Administration of Hospital Clinical Medicine Development Special Funding (ZYLX201511)** supported this work; it is available for open access with attribution under terms of the Creative Commons Attribution 4.0 International (CC BY 4.0).

3.2 Data Preprocessing

To facilitate effective data loading and augmentation, the dataset was organized as a hierarchical directory structure. To ensure compatibility with Keras' ImageDataGenerator for simple batch generation during training and testing, each subfolder contained images that belonged to a single class label.

Data generators were implemented using the TensorFlow/Keras libraries. The generators were allowed to Both the training and test datasets' pixel intensities are rescaled to the $[0,1]$ interval, the training dataset is improved in real time, and memory consumption is decreased through real-time batch loading. When it was noted that annotated medical imaging datasets were smaller than natural image datasets, strong augmentation was employed to enhance the model's generalization [28].

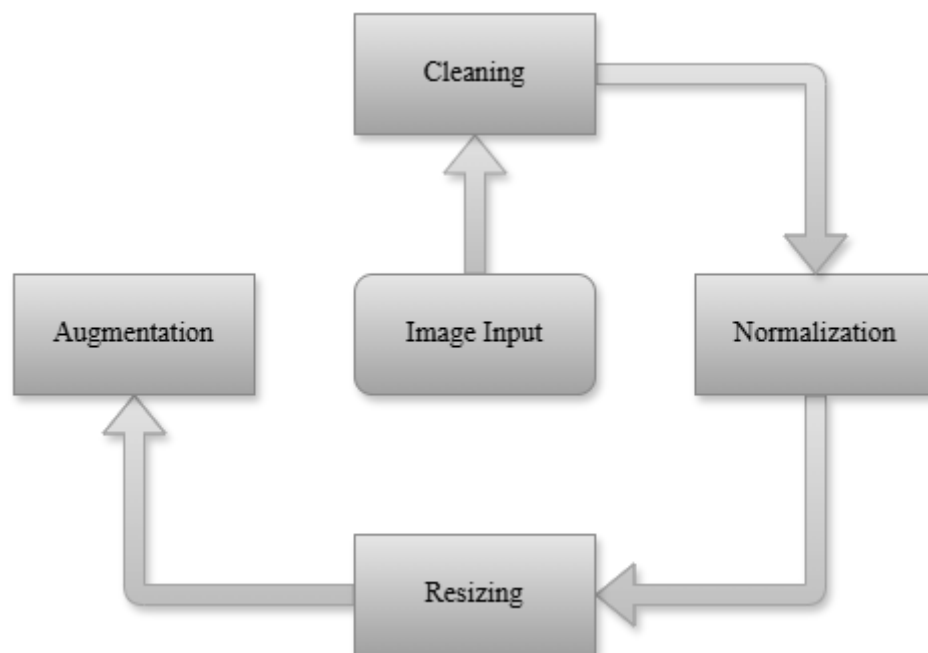


Figure 3.1: Data Pre-processing

3.4.1 Data Cleaning

Cleaning the LUNG-PET-CT-DX dataset was an essential preprocessing step in getting it ready for model training. The entire process ensured that disparities were eliminated and generally improved the quality of inputs to deep learning, which was necessary given the type and volume of medical imaging data.

To avoid training errors, non-image files and corrupted DICOM instances were eliminated. Pictures with inadequate labels or metadata were disqualified. To guarantee correspondence and label consistency, the PET and CT image pairs were cross-checked. Additionally, the images were standardized to a single format (like PNG or JPEG) and examined for anomalous conditions, like empty slices, excessive brightness/contrast artifacts, and manual cropping of non-lung views.

3.4.2 Resizing and Normalization

All the images that were downsized to a fixed resolution of 512x512 pixels (Resizing). Resizing images to a standard size assists with batch processing and allows the use of pretrained models, which frequently require their input sizes to be fixed (e.g., 512x512 for ResNet models) [20]. Pixel values normalized to [0,1] (Normalization). Normalization of pixel intensity values from 0 to 1 enhances the training convergence and maintains the numerical stability of the model [21].

3.4.3 Augmentation Techniques

Several augmentation techniques were performed to simulate realistic variations (e.g. patient positioning, differences in exposure) without breaking clinical realism.

Rotation involves rotating the image around its center by a random angle within a specified range (e.g., $\pm 15^\circ$). This augmentation helps the model become invariant to the orientation of structures, which is particularly important because small changes in patient positioning are common in medical imaging.

Mathematical Representation:

For an image point (x, y) rotated by an angle θ around the origin, the new coordinates (x', y') are [29]:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \quad (1)$$

Where:

θ is the rotation angle (positive for counter-clockwise rotation).

Translation shifts image along the X (horizontal) or Y (vertical) axis by a certain proportion (e.g., 10% of width or height). This augmentation simulates patient movement or slight differences in scan alignment.

Mathematical Representation:

A translation by t_x and t_y is given by [30]:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (2)$$

Where:

- t_x and t_y are translations along X and Y.

Shearing is a geometric transformation that tilts the image along the X or Y axis. It distorts the image such that the shape appears "slanted," which helps the model learn features invariant to shape deformation.

Mathematical Representation:

For a horizontal shear [31]:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} 1 & s \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \quad (3)$$

Where:

- s is the shear factor.

Zooming either enlarges (zoom in) or shrinks (zoom out) the image within a defined range (e.g., 90–110%). Zooming augments scale variation, helping models recognize objects at different scales.

Mathematical Representation:

Zooming can be treated as a scaling transformation:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} s_x & 0 \\ 0 & s_y \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \quad (4)$$

Where:

- s_x and s_y are scaling factors along the X and Y axes [32].

If $s_x, s_y > 1$: Zoom in

If $s_x, s_y < 1$: Zoom out

Brightness Adjustment changes the intensity of the image pixels by either increasing or decreasing brightness, simulating variations in lighting conditions or scanner calibrations.

Mathematical Representation:

Brightness adjustment modifies each pixel $I(x, y)$ according to [33]:

$$I'(x, y) = I(x, y) + \Delta B \quad (5)$$

Where:

- ΔB is the random brightness offset value.

This value is usually selected randomly within a small range, e.g., $\pm 20\%$.

Contrast adjustment alters pixel intensity differences. A higher contrast accentuates differences between darker and lighter areas, but a lower contrast brings them closer to a similarity.

Mathematical Representation:

Contrast adjustment is applied as [34]:

$$I'(x, y) = \alpha \times (I(x, y) - \mu) + \mu \quad (6)$$

Where:

- α is the contrast factor (>1 increases contrast, <1 decreases contrast),
- μ is the mean pixel intensity of the image.

Higher contrast highlights edges and boundaries, helping the model better learn shape-related features.

Gaussian noise is artificially added to simulate scanner noise, motion artifacts, or electronic fluctuations, making the model robust to real-world clinical imaging conditions.

Mathematical Representation:

Gaussian noise is added as [35]:

$$I'(x, y) = I(x, y) + \mathcal{N}(0, \sigma^2) \quad (7)$$

Where:

- $\mathcal{N}(0, \sigma^2)$ is a Gaussian distribution with mean 0 and variance σ^2 .

Low values of σ (standard deviation) are used to ensure subtle, realistic noise without overwhelming the original signal.

All augmentations applied:

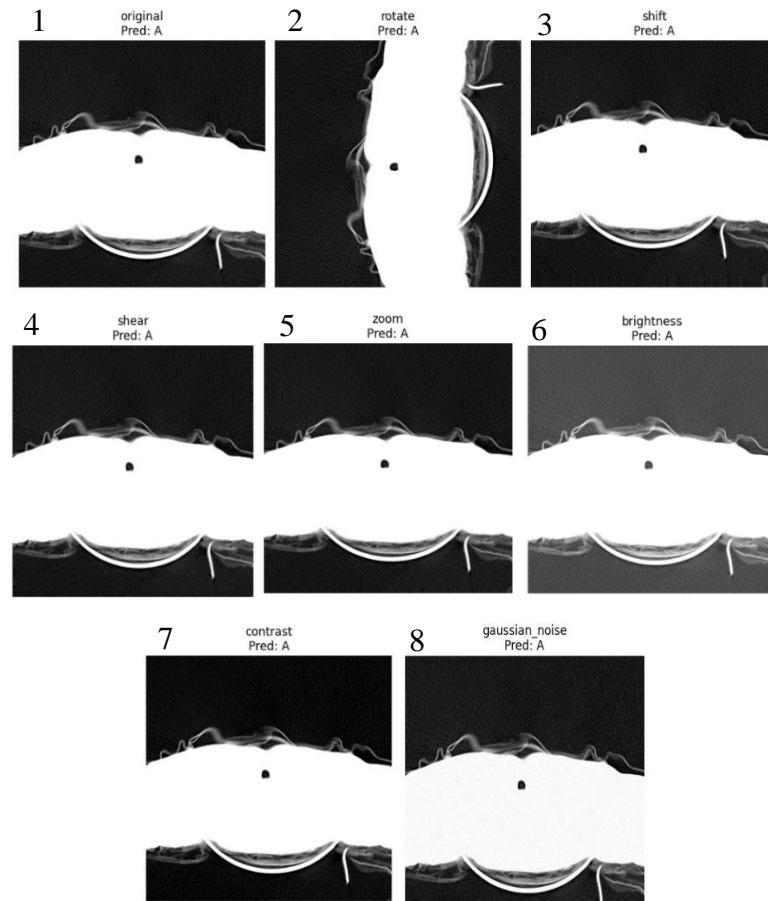


Image **1** is original, rotation applied (**2**), shift (**3**), shear (**4**), zoom (**5**), brightness (**6**), contrast (**7**), Gaussian noise (**8**).

Taking Medical imaging into account heavily, horizontal and vertical flipping were not applied at augmentation time. In medical imaging, anatomical variations (such as the distinct characteristics of the right and left lungs) present critical diagnostic information. Reversing the images may result in unrealistic anatomical models and negatively impact model learning [19].

No augmentations were applied to the test dataset so that the scores in the evaluation would accurately indicate model performance on unadulterated, real data. Only basic pre-processing operations were carried out.

Training was performed with a batch size of 32. This batch size was chosen to balance memory efficiency against stability in gradient estimation. The final input tensor shape for the models was thus (32, 512, 512, 3), which corresponds to batches of 32 RGB images resized to 512x512 pixels. These enhancements enhance the effective size of the training set and assist in avoiding overfitting on the scarce medical data.

3.5 Feature Extraction and Fine tuning

In order to utilize the advantages of transfer learning and enhance model generalization on the LUNG-PET-CT-DX dataset [13], a two-stage training approach was employed: feature extraction and fine-tuning.

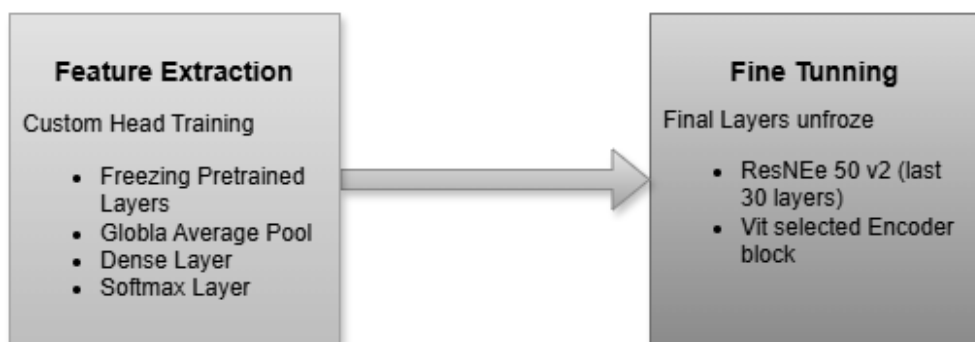


Figure 3.2: Feature Extraction and Fine Tuning

In the initial phase, the base model ImageNet-pretrained ResNet50V2 was used as a fixed feature extractor. The weights of all the convolutional layers of the ResNet50V2 backbone were not updated during training, (i.e they were frozen). A custom classifier head was added on top of the frozen layers, which consisted of Batch Normalization, Global Average Pooling, and a Dense layer with a softmax activation function for multi-class classification. Training here was focused only on the recently added layers. This allowed the model to extrapolate the acquired high-level features to the specific classes of lung diseases in the dataset without changing the pretrained representations, thereby reducing the risk of overfitting on the scarce medical data.

Only the parameters θ_{head} are updated:

$$\theta_{\text{head}}^* = \arg \min_{\theta_{\text{head}}} \mathcal{L}(\theta_{\text{head}}; X, y) \quad (8)$$

where \mathcal{L} is the categorical cross-entropy loss. Training is performed for 20 epochs with a relatively high learning rate (e.g., 1×10^{-3}).

When the head of classification stabilized and performed well enough, then the model was changed to fine-tuning mode. In this case, the top 30 layers of the ResNet50V2 backbone were not frozen, but instead were left free to enable selective update of higher-level convolutional features by ongoing training. To avoid the pretrained weights from being disrupted heavily, a lower learning rate (i.e., $1e-5$) was used along with the Cosine Decay with Restarts scheduler. This fine-tuning gradually allowed the model to adapt deeper representations to the medical imaging domain, improving class separability and overall performance. Fine-tuning is conducted using a significantly small learning rate (e.g., 1×10^{-5}) to prevent catastrophic forgetting:

$$\theta^* = \arg \min_{\theta} \mathcal{L}(\theta; X, y), \quad \text{where } \theta = \{\theta_{\text{base}}, \theta_{\text{head}}\} \quad (9)$$

This hierarchical training strategy enables the network to maintain its pre-learned representations while gradually specializing for the lung cancer classification task.

3.6 Model Architecture

Our study presents a hybrid deep learning model that synergistically integrates the spatial feature extraction strength of Convolutional Neural Networks (CNNs) and the global context modelling strength of Vision Transformers (ViTs). The model is carefully crafted to improve classification accuracy on the LUNG-PET-CT-DX dataset using both anatomical and functional imaging information.

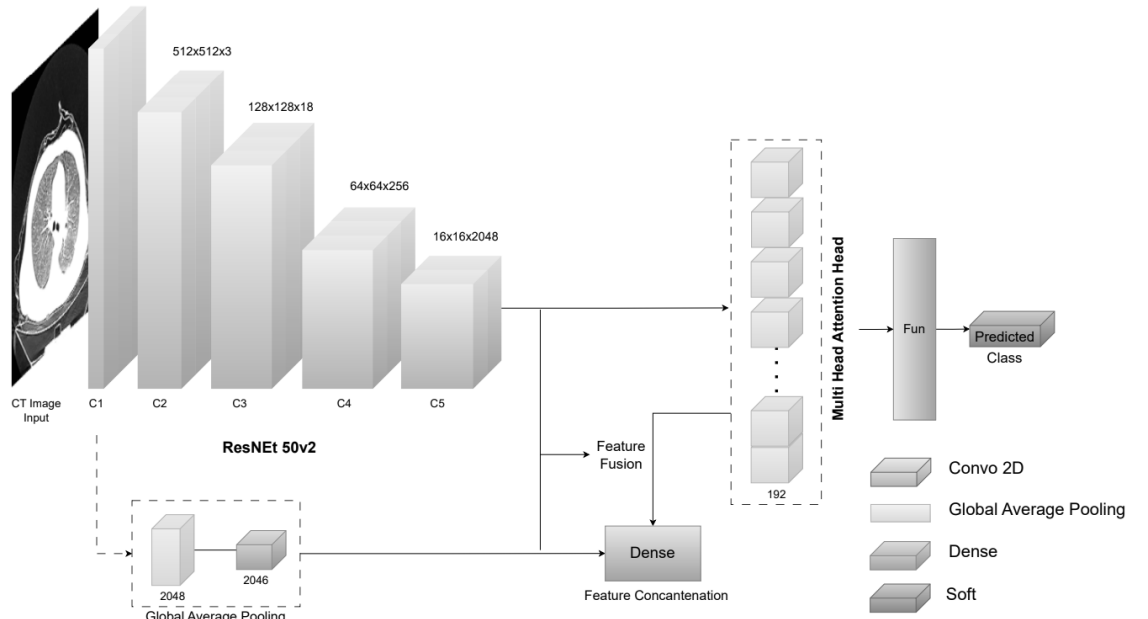


Figure 3.3: Model Architecture

3.6.1 Base Model: ResNet50V2

The baseline of the intended architecture is the ResNet50V2 model, which is famous for its deep residual learning mechanism that resists the vanishing gradient problem in deep networks. By starting ResNet50V2 from weights pretrained on the ImageNet dataset, the model enjoys richer feature representations obtained from a huge body of images, which in turn allows it to converge better and generalize on medical imaging problems.

In order to adapt the pretrained ResNet50V2 model to the particular needs of the LUNG-PET-CT-DX dataset, the standard fully connected classifier head is discarded. This adaptation enables the addition of custom layers that are better suited to the intricacies of multimodal medical imaging data. The model is trained to take resized input images to 512x512 pixels having three color channels, which coincides with ResNet50V2's input size expectations. In order not to lose valuable prelearned low-level features, all ResNet50V2 base layers are frozen through initial training phases. This is done to lock down the grounding representations while having the later layers learn the specificities of the target dataset.

3.6.2 Custom Top Layers

Following the base model, a Batch Normalization layer is added to normalize the inputs to the next layers. This normalization speeds up training, improves stability, and lessens sensitivity to network initialization, thus enhancing overall model performance.

A Global Average Pooling (GAP) layer is used to condense the spatial feature maps into a reduced, fixed-size feature vector. GAP minimizes the overall number of parameters of the model, preventing overfitting hazards and enabling seamless fusion with transformer-based parts. To connect the CNN-features with the transformer module, the output from the GAP layer is reformatted into a sequence format suitable for transformer architecture. This conversion allows the model to access the data as a sequence of tokens, where each token represents different features of the input image, seamlessly connecting the CNN and transformer modules.

3.6.3 Vision Transformer Attention Block

The ViTAttention block introduces the Multi-Head Attention mechanism, enabling the model to pay attention to information from multiple representation subspaces simultaneously. This mechanism facilitates the model to better capture subtle relationships and dependencies in the data, which is especially useful for difficult medical imaging tasks.

To stabilize and optimize the training of the transformer module, Layer Normalization is used prior to each sub-layer and residual connections are made around each sub-layer. These architectural design decisions enable improved gradient flow, faster convergence, and enhanced generalization through the retention of the original input information and avoidance of the vanishing gradient risk. The transformer module is set with an embedding size of 512 and 8 attention heads. This setting finds a compromise between model complexity and computational complexity, allowing the model to identify varied patterns and features in the data without requiring excessive computational expenses.

Combining transformer-based attention mechanisms with the CNN architecture gives the model the capability to learn both local and global features efficiently. Although CNNs are good at learning localized patterns, transformers are capable of modeling long-range dependencies and global context. This combination utilizes the strengths of both architectures and results in a better understanding of the input data and, therefore, better classification performance.

3.6.4 Output Layer: 4-Class Softmax Activation

The last layer of the model is a Dense layer with Softmax activation function, which outputs probability distributions over the four target classes. Such a configuration is useful for multi-class classification because it allows the model to assign probabilities to each class, hence enabling fine-grained decision-making in clinical applications.

3.7 Training Strategy

The training pipeline carried out in our project is intended to enable maximum performance from two independent but complementary deep-learning architectures: ResNet50V2-a convolutional neural network (CNN) architecture known for its residual learning-and Vision Transformer (ViT)-a transformer-based model designed specifically for image understanding through self-attention mechanisms. In the present work, our models were trained using a two-phase approach to transfer learning.

Feature extraction was carried out with a ResNet50V2 backbone pretrained on ImageNet; all layers except the convolutional ones were frozen to serve as a fixed feature extractor `keras.io`. Initially, only the newly added classification layers were allowed to train. This stage enabled the network to learn the mapping from rich ImageNet features to classes of lung disease. Once the added layers were trained, we entered the fine-tuning phase, during which the deepest layers of ResNet50V2 were unfrozen. In particular, the last 30 convolutional layers were made trainable, allowing the model to modify its high-level feature representations for the lung-imaging realm. This two-step procedure—one wherein a pretrained CNN is first frozen and then selectively unfrozen—is a time-honored best practice in transfer learning and has been shown to yield good performance with limited target data.

ResNet50V2 (an upgraded version of ResNet with identity mappings) was instantiated without its top classification head `kn`. The weights of the network were initialized from ImageNet training, and all layers were non-trainable at first. A new global pooling and dense classifier was appended. At this point, the ResNet50V2 weights provided a fixed feature extractor, and only the last classifier layers were trained on the lung CT dataset. This approach retains the power of low-level features learned on the large dataset while adapting high-level decision functions to the medical task. Feature extraction this way takes advantage of ResNet50V2's vast representational ability without compromising the small lung dataset. Fine-Tuning: It was only after the classifier stabilized that layers were unfrozen, the last 30 layers of the backbone being ResNet50V2. This led to higher-level features being fine-tuned for lung-related imaging. Low learning rates were employed during fine-tuning for adapting these pretrained filters. This unfreezing is selective and therefore able to balance the demands for adaptation with the risk of overfitting: It gives the model the potential to adapt very abstract features to medical patterns while keeping earlier layers intact with visual filters. Such layer-wise fine-tuning is known to yield superior generalization compared to training from scratch, especially in specialized domains, and therefore is worth the effort.

The model uses its own ViTAttention module to account for global context and long-range dependencies. The diagram below follows the Vision Transformer (ViT) paradigm: The input image (or feature map) is partitioned into a sequence of flattened

patches, linearly embedded, and augmented with positional encodings (see [ar5iv.org](https://arxiv.org), en.wikipedia.org). A learnable classification token is prepended to this patch sequence. The sequence is then passed through the standard Transformer encoder forming a series of blocks, alternating between multi-headed self-attention (MSA) layers and feed-forward (MLP) blocks, together with layer normalization and residual connections (see [ar5iv.org](https://arxiv.org)). The MSA constitutes a mechanism by which all the patch tokens can attend to each other, thus encoding global information. After the Transformer layers, the output corresponding to the class token will be used for classification.

The ViTAttention layer was used as a Keras custom layer with 8 attention heads and an embedding dimension of 128 (equal to the ResNet feature dimensionality). Each attention block performs scaled dot-product attention in parallel across heads, followed by a feed-forward network and skip connections, similar to the original Transformer design [ar5iv.org](https://arxiv.org). This block was placed following the ResNet50V2 feature extractor (i.e., the ResNet output was reshaped as a sequence of 128-dimensional patch embeddings) so that both local convolutional features and global self-attention are employed. This hybrid CNN+Transformer combination goes beyond standard pipelines by bringing ResNet's locality together with ViT's global context modeling, which is especially useful for intricate medical images where subtle patterns could be spatially far apart.

At training time, the ViTAttention block was trained jointly with the remainder of the network. Due to the fact that the lung dataset is small, we did not pretrain the Transformer block over a huge external corpus but rather initialized it randomly and trained it end-to-end on the LUNG-PET-CT-DX data. Nevertheless, applying positional encodings and the class token adheres to standard ViT practice [ar5iv.org](https://arxiv.org)[ar5iv.org](https://arxiv.org). This precise engineering provides a mechanism whereby the attention module may attend to diagnostically significant areas throughout the entire image.

Table 3.1: Hyperparameters

Hyperparameter	Value / Setting
Batch Size	32
Initial Learning Rate	0.001 (feature extraction) 0.0001 (fine-tuning)
Optimizer	Adamax
Learning Rate Scheduler	Cosine Decay with Restarts
Base Model	ResNet50V2 (pretrained on ImageNet)
Fine-Tuned Layers	Last 30 layers of ResNet50V2
Transformer Block	Custom ViTAttention block
Attention Heads	8
Embedding Dimension	128
Loss Function	Categorical Crossentropy
Dropout Rate	0.3
L2 Regularization	0.001
Epoch	20

3.7.1 Optimizer and Learning Rate Scheduler

3.7.1.1 Cosine Decay with Warm Restarts

Training employed the Adamax optimizer, an adaptive first-order optimizer in terms of the infinity norm `keras.io`. Adamax was selected because of its insensitivity to the difference in gradient scales and because it can learn multiple learning rates for each parameter, which is helpful in noisy or complicated tasks. The standard hyperparameters ($\beta_1=0.9$, $\beta_2=0.999$) were used. For stability in optimization, we employed an initial learning rate of $1e-3$ for the training in the frozen stage and a lower rate (e.g. $1e-4$) in fine-tuning. For improving model convergence and reducing suboptimal local minima, we employed the Stochastic Gradient Descent with Warm Restarts (SGDR) according to the Cosine Decay with Restarts (CDR) scheduling

method introduced by Loshchilov and Hutter [36]. The training step t learning rate is corrected according to a cosine function:

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min}) \left(1 + \cos \left(\frac{T_{cur}}{T_i} \pi \right) \right) \quad (10)$$

where:

- η_{min} is the minimum learning rate.
- η_{max} the initial learning rate before decay.
- T_{cur} is the number of epochs since the last restart.
- T_i is the number of epochs before the next restart.

Each restart allows the learning rate to increase back to η_{max} , helping the model escape saddle points and shallow local minima. Following parameters were considered while implementing.

- **initial_learning_rate**: Set to 1×10^{-3} for feature extraction and lowered to 1×10^{-5} during fine-tuning,
- **first_decay_steps**: Defines the interval before the first restart,
- **t_mul** and **m_mul**: Multipliers for successive decay steps and amplitude reduction,
- **alpha**: Controls the minimum learning rate fraction, typically $\alpha=0.0$.

This technique has been empirically shown to improve generalization across various architectures and tasks [36].

3.7.1.2 Callback Strategies

To save the model state that achieves the best validation performance, the ModelCheckpoint callback is used. This tracks a given metric usually validation loss and saves the weights only when improvement is noted. This minimizes the risk of overfitting and model deterioration during training epochs.

Let $L_{val}^{(i)}$ be the validation loss at epoch i . The weights θ are saved only if:

$$L_{val}^{(i)} < \min_{j < i} L_{val}^{(j)} \quad (11)$$

Checkpointing ensures retention of the most performant model for later evaluation.

To avoid unnecessary training after convergence is observed, EarlyStopping is optionally used. This callback stops training if no validation loss improvement is observed within a given "patience" window, typically 5–10 epochs. It imposes an early stopping condition:

Stop Training if $L_{\text{val}}^{(i)} \geq \min_{j < i} L_{\text{val}}^{(j)}$ for p consecutive epochs

This strategy ensures computational efficiency and model generalization [37].

Two-Stage Training Procedure

The training procedure is decomposed into two separate stages to fully leverage transfer learning while permitting fine-tuning on task-specific data.

Stage 1: Feature Extraction

In this early phase, the convolutional backbone (ResNet50V2) and the transformer encoder (ViT) are frozen, and only the custom classification head is trained. This head generally contains:

- Global Average Pooling (or Flattening for ViT),
- Dense layers with dropout regularization,
- Final Softmax output layer for multi-class classification.

Only the parameters θ_{head} are updated:

$$\theta_{\text{head}}^* = \arg \min_{\theta_{\text{head}}} \mathcal{L}(\theta_{\text{head}}; X, y) \quad (12)$$

where \mathcal{L} is the categorical cross-entropy loss. Training is performed for 20 epochs with a relatively high learning rate (e.g., 1×10^{-3}).

Stage 2: Fine-Tuning

After the head has sufficiently adapted to the task, the last layers of the base model are unfrozen. In the case of ResNet50V2, the top 30 layers are made trainable so that high-level semantic features can be trained on the medical imaging space. In the case of ViT, chosen transformer encoder blocks are unfrozen similarly. Fine-tuning

is conducted using a significantly small learning rate (e.g., 1×10^{-5}) to prevent catastrophic forgetting:

$$\theta^* = \arg \min_{\theta} \mathcal{L}(\theta; X, y), \quad \text{where } \theta = \{\theta_{\text{base}}, \theta_{\text{head}}\} \quad (13)$$

This hierarchical training strategy enables the network to maintain its pre-learned representations while gradually specializing for the lung cancer classification task.

3.8 Evaluation Metrics

In order to thoroughly evaluate the performance of the suggested hybrid ResNet50V2-ViTAttention model for lung cancer classification, various conventional evaluation measures were utilized: accuracy, precision, recall, F1-score, and Area Under the ROC Curve (AUC). They provide information about various aspects of model performance, particularly important in medical diagnosis where both false negatives and false positives have serious consequences.

3.8.1 Accuracy

Accuracy measures overall model correctness by measuring the percentage of correctly classified samples:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (14)$$

where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

Accuracy works best when class distribution is even.

3.8.2 Precision

Precision measures the model's capability to correctly identify only relevant (positive) instances:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (15)$$

High precision means low false positives, which is crucial in the healthcare sector to prevent misdiagnosis of healthy patients.

3.8.3 3. Recall (Sensitivity)

Recall measures the model's ability to detect all relevant instances (i.e., all true positives):

$$\text{Recall} = \frac{TP}{TP+FN} \quad (16)$$

High recall is essential to prevent diseased patients from being missed.

3.8.4 F1-Score

F1-score offers a harmonic mean of precision and recall, trading off both metrics:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (17)$$

It is especially helpful in imbalanced classification situations where a single metric may not be sufficient.

3.8.5 ROC Curve and AUC (Area Under Curve)

The Receiver Operating Characteristic (ROC) curve graphs the True Positive Rate (Recall) against the False Positive Rate:

$$\text{FPR} = \frac{FP}{FP+TN} \quad (18)$$

The AUC is the probability that the model will prefer a randomly selected positive instance over a randomly selected negative instance. Values of AUC near 1.0 signal good model performance.

3.9 Chatbot Module

In order to enhance further the interpretability and educational value of the proposed lung disease classification system, a separate chatbot module has been designed and integrated into the system. The chatbot is a smart question answering aid that is designed specifically for dealing with user queries related to human lung anatomy. Its integration follows a Retrieval Augmented Generation (RAG) architecture for presenting domain-specific, evidence-based answers. Unlike other chatbots that can only reply in pre-determined answers or stock language models, the RAG approach provides grounded answers based on factual information in a specially curated knowledge base of lung anatomy.

3.9.1 Knowledge Source and Data Preparation

The medical domain knowledge of the chatbot comes from a medically reviewed set of PDFs on comprehensive lung anatomy including features such as alveoli, bronchi, lobes, and pleura. These files were parsed and cleaned to produce high-quality, paragraph-level text chunks. Each chunk of text was then inserted into a

high-dimensional vector space through a sentence-transformer model, which maintains the semantic similarity of anatomical concepts. These embeddings were indexed with Facebook AI Similarity Search (FAISS), a fast vector database that is optimized for nearest neighbour search operations.

This preparation phase ensures that the chatbot is fed with structured, contextual information on the human respiratory system so that it can generate specific and relevant answers to user questions.

3.9.2 Query Processing and Retrieval Pipeline

Upon a user inputting a lung-related question through the chatbot interface, the backend invokes a multi-stage pipeline to retrieve and generate the response. The input question submitted is first pre-processed with simple pre-processing like keyword extraction and sentence embedding based on the same embedding model employed at the data preparation phase. The query embedding is then used to perform a top-k similarity search in the FAISS index, retrieving the most semantically similar anatomical paragraphs.

These recovered contexts are incorporated into the user's original question to form an enriched prompt. This combined input is essential in order to ground the chatbot's response on fact-based knowledge rather than relying solely on the model's internal parameters.

3.9.3 Response Generation using Gamma API

The improved prompt is passed on to a hosted large language model (LLM) using the Gamma API, which carries out the generation of the final output. The LLM is designed to generate natural language answers that are understandable, medically accurate, and relevant to the original user query. Because the model is restricted by the accessed anatomical context, the resulting answers have factual integrity and interpretability, especially important in a medical setting.

The final response is returned to the frontend as JSON and output in the chatbot window in real time. The end-to-end cycle from query submission to display of response is optimized to be fast, light, and suitable for low-latency interactions.

3.9.4 Design Benefits and Considerations

This chatbot design has several advantages within a clinical or learning setting:

- **Domain-Specific Accuracy:** Use of an embedded lung anatomy dataset ensures that answers are relevant and medically contextualized.
- **Explain ability:** Since answers are computed from pulled factual content, they can be traced back to source documents.
- **Low Latency:** Use of FAISS for vector search and asynchronous API requests ensures fast response times.
- **Scalability:** The architecture allows for future expansion to other anatomical spaces (e.g., heart, liver) by changing the dataset and re-indexing with FAISS.

3.9.5 Integration with the Main System

The chatbot is deployed as a microservice in the existing FastAPI backend. It is executed on a different endpoint and interacts with the frontend asynchronously. The modular architecture allows the chatbot to be executed independently of the classification model, but share the same user interface and authentication mechanisms. This integration enables users to engage with the system in a multi-modal fashion using image based diagnostics alongside text-based anatomical questioning.

3.10 Chapter Summary

This chapter describes the intricate technological underpinnings of the novel system designed for lung disease classification and anatomical assistance. The LUNG-PET-CT-DX dataset was developed to effectively manage the situation, which included the details of 250,000 and over 355 patients who were confirmed to have lung cancer via histopathology. Additionally, the dataset comprised multimodal PET and CT data. However, the dataset was cleaned and resized, and 512×512 pixels were normalized, and medically healthy augmentation methods (e.g., rotation, translation, brightness, and noise addition) were used, while anatomical distortion strategies like flipping were avoided.

The system incorporates a two-stage training strategy. Feature extraction is done using a ResNet50V2 backbone whose layers are frozen. Later the 30 layers from the top are unfrozen, and the lower learning rate is used with a Cosine Decay with Warm Restarts scheduler and Adamax optimizer. A Vision Transformer (ViTAttention) block comes after the ResNet50V2 feature extractor to seize global context and long-range dependencies. The soup of the CNN–ViT model topped with the softmax output layer provides four-class lung disease classification. Apart from the prediction being made based on images, the system was also able to include a chatbot module for the purpose of lung anatomy question answering. Using a FAISS vector search and a large language model served via the Gamma API for the Retrieve-Augmented Generation (RAG) architecture, an effort is made to imitate the chatbot; it is natural that both sides understand each other. The user's query is first embedded and then matched pragmatically against a pre-included lung anatomy PDF dataset. The context that is suited to the problem is derived and is then added to and elevated the search, and finally, the long language model (LLM) is in charge of producing the correct medical and actionable response.

The machine guarantees to not only provide high-level performance , explainability (via LIME) , less latency, but also modularization. It is presented by means of a web-based application that is made up of different technologies such as MERN stack and FastAPI. This enables the features like real-time prediction, image upload, and asking of the anatomical data. The System Architecture diagram provided below present a visual insight of pipeline for both Diagnostic and chatbot flow.

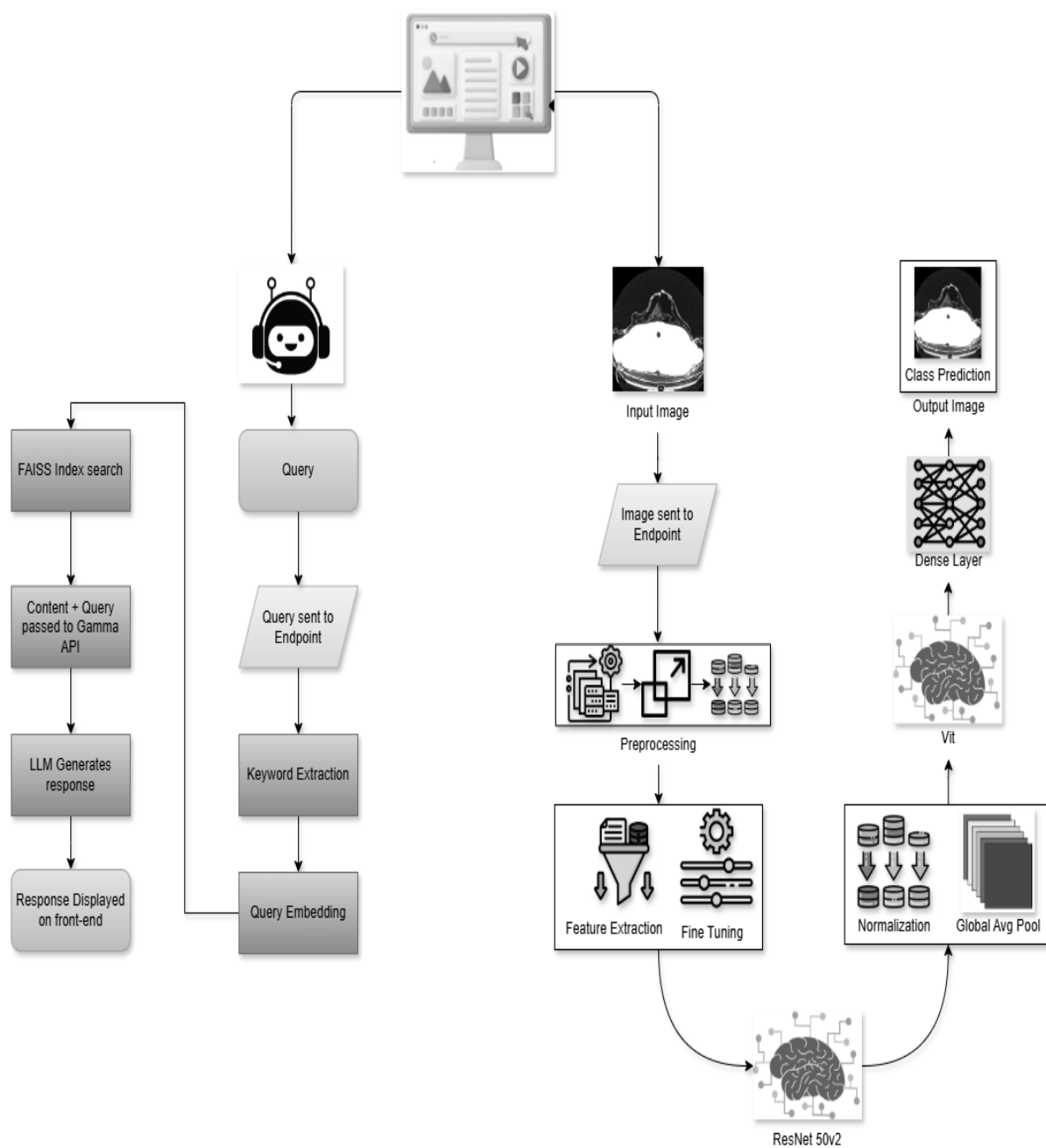


Figure 3.4: System Architecture

The system architecture diagram shows how the developed platform is able to perform two functions at once, thus making the process of both the lung disease classification pipeline and the lung anatomy chatbot module clear for understanding. The image classification flow starts with a PET/CT scan, that the user uploads through the frontend, on the right side. The image is sent to the backend, where some preprocessing steps are executed (resizing, normalization), and then a hybrid model is used for inference (ResNet50V2 and ViTAttention). The model predicts the disease class and the model's confidence returned and displayed in the user interface.

While in the left part, the chatbot flow begins when the user submits a question that is related to the lung. Then, the back end extracts keywords from the question, embeds the query, and checks a FAISS index built from a standardized lung anatomy dataset. The answer with the context is passed to the language model via the Gamma API, which helps in creating a perfectly focused response. Finally, the response is sent back to the user through the frontend to be displayed in the chatbot window.

CHAPTER 4

IMPLEMENTATION

4.1 System Workflow Overview

The system, as devised, offers a holistic solution for automated lung disease detection and anatomical assistance. To begin, the user uploads a PET/CT lung image through a web interface. The image is preprocessed on the backend before being fed to a hybrid deep learning model made up of ResNet50V2 and Vision Transformer (ViTAttention). The model returns a predicted disease class output with a score of confidence, which is presented to the end user. In addition to disease classification, there is also a chatbot integrated into the interface that responds to a user's inquiries related to lung anatomy. This chatbot operates using a Retrieval-Augmented Generation (RAG) methodology to generate relevant, medically accurate, and explainable answers.

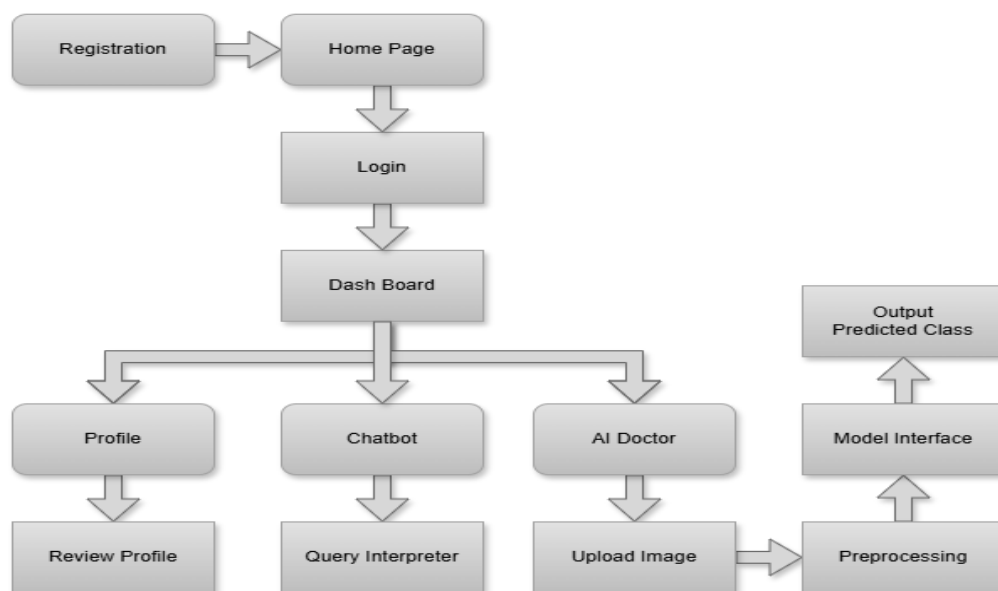


Figure 4.1: Web interface Flow

In other words, the entire process is made to be efficient and accessible, enabling clinicians or users at the same time to receive diagnostic predictions and anatomical support on the go, within a very unified platform. Image input, followed by data preprocessing, model inference, query input by the user, and results streaming, is the flow of the entire process.

4.2 Technologies Used

The system makes use of a heterogeneous and well-orchestrated technology stack, ranging from machine learning, web development, to deployment.

4.2.1 Machine Learning and Modeling

For the machine learning aspect, the model was developed with TensorFlow and Keras. The architecture combines a ResNet50V2 convolutional backbone for spatial feature extraction and a Vision Transformer block to identify long-range dependencies and global context. This hybrid setup enhances classification accuracy and robustness across heterogeneous imaging conditions.

4.2.2 Chatbot System

The chatbot is based on a Retrieval-Augmented Generation (RAG) model. Domain-specific lung anatomy information is gathered from expert-reviewed PDFs, which are parsed and embedded by a sentence-transformer model. The generated embeddings are indexed in a FAISS (Facebook AI Similarity Search) vector database. When a user enters a query, the system retrieves the top-k most similar contexts from FAISS based on the relevant keywords. These are subsequently sent to a big language model through the Gamma API, and it comes back with a response based on what it fetched.

4.2.3 Web Application Stack

The frontend is coded in React (MERN stack), offering a user-friendly and responsive user interface. The backend is done with FastAPI, selected for its performance and async nature. User data is handled by MongoDB, with optional storage of chat history or image processing records. The whole platform is hosted on a cloud hosting provider to provide public access and maintain uptime and scalability.

4.3 Web Interface Features

The web interface is built to give users an effortless diagnostic and information experience. PET/CT images can be uploaded directly via the browser. The uploaded image is checked and sent securely to the backend, where preprocessing operations like resizing to 512x512 pixels and normalization are performed. The processed image is then sent to the AI model, and the classification output is returned with a probability score.

Users can also interact with the lung anatomy chatbot. This functionality is integrated into the interface as a chat window, where users can enter questions. The chatbot answers these questions in real time, searches the vector database for context, and returns LLM-generated responses.

This enables users not only to get a disease prediction but also to gain additional information about associated anatomical structures and functions, enhancing interpretability and clinical utility. Other features include secure login and registration capabilities. These enhance user personalization without compromising system performance.

4.4 Integration with Backend Systems

The backend is the running core of the application. When started, the trained hybrid model is deserialized into memory from a TensorFlow SavedModel or HDF5 format. FastAPI offers endpoint interfaces that receive image submission, query reception, and response transmission.

When an image is uploaded, the backend does preprocessing like format change, resizing, and pixel scaling. It then passes the image to the classification model, which outputs the category of the disease among four given classes. The output is the predicted label along with its confidence, presented in the form of a JSON object and sent back to the frontend to be displayed.

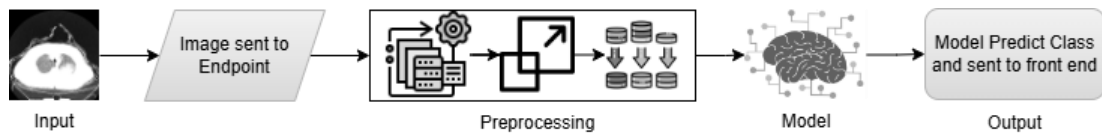


Figure 4.2: Pipeline of Diagnosis System

For chatbot questions, the backend performs keyword extraction and embedding with the same model applied to document embedding. The FAISS index is searched for the most contextually relevant text segments, which are appended to the user's question and passed to the Gamma API. The language model generates a full response based on both the query and context retrieved, making sure the answer is medically pertinent and grounded in the lung anatomy dataset.

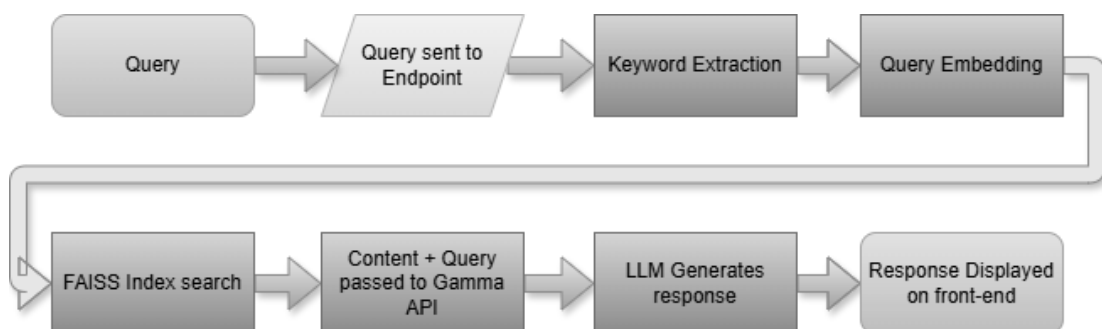
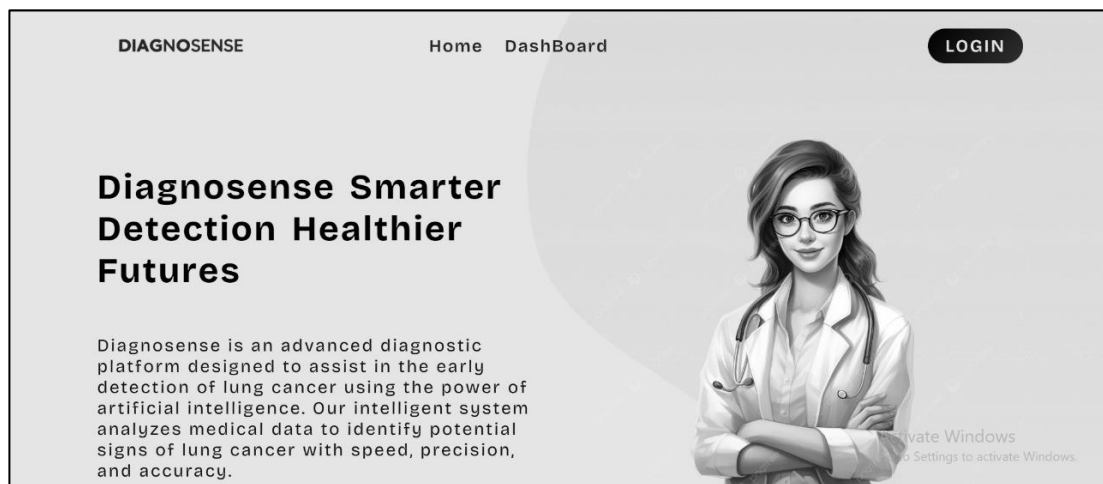


Figure 4.3: Pipeline of Chatbot

These elements operate asynchronously for low-latency support. JSON is employed in data exchange, allowing a lightweight and uniform pattern of communication between the frontend and backend modules.

4.5 Visual Representation of Flow

Considering robust and easy interaction, user friendly interface is developed maintaining simple flow. First a landing page (**Home Page**) is being utilized for presenting necessary information.



- Top menu bar contains 3 buttons **Home** to return back to Home, **Dashboard** to access features and **Login** button. For accessing dashboard registration is **mandatory**.

- Upon registration, credentials are stored employing hashed credentials coupled with JWT-based session management to protect user logins. User profile data

is securely stored in MongoDB with access controls imposed on protected routes. Once registered Login to access dashboard

DIAGNOSENSE Home Dashboard LOGIN

Sign In

Please Login To Continue

Email

Password

Confirm Password

Not Registered? Register Now

Login

DIAGNOSENSE Quick Links
Home
Dashboard

Developed By
Muhammad
Yasin
Muzammal Bilal
Muhammad
Shahzaib

Activate Windows
Go to Settings to activate Windows.

- After successful login, Further options will appear. Keeping a clear and precise outlook, features are visible and easy to interact. User Profile will be visible displaying all necessary information.

DIAGNOSENSE

Dashboard Profile AI Doctor ChatBot

User Profile

View and manage your account details

First Name Last Name
ahmad bilal

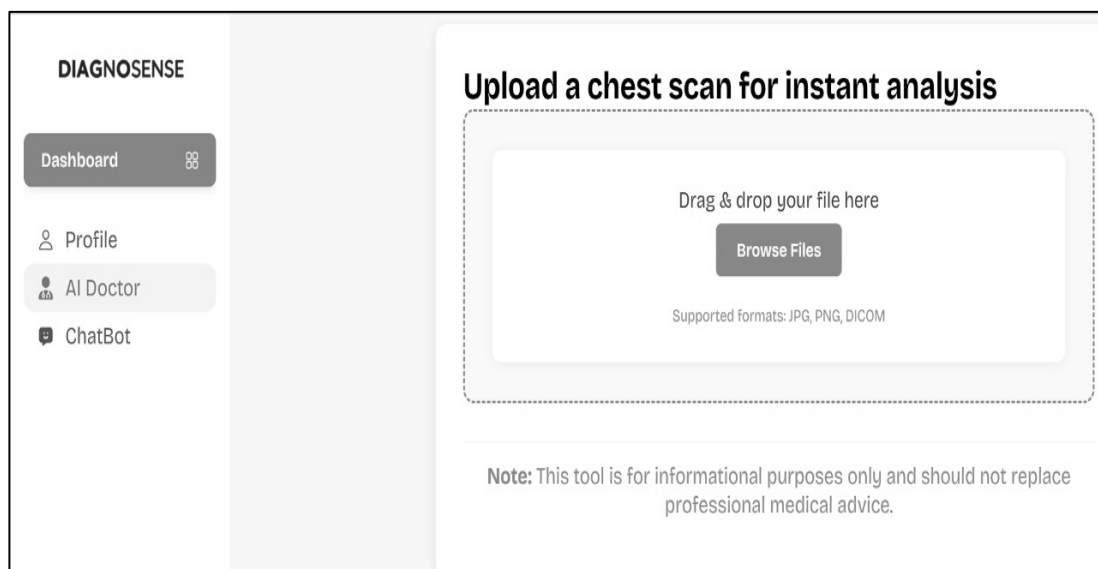
Email Phone
m@gmail.com 03174849258

NIC Date of Birth
3520222647842 5/1/2025

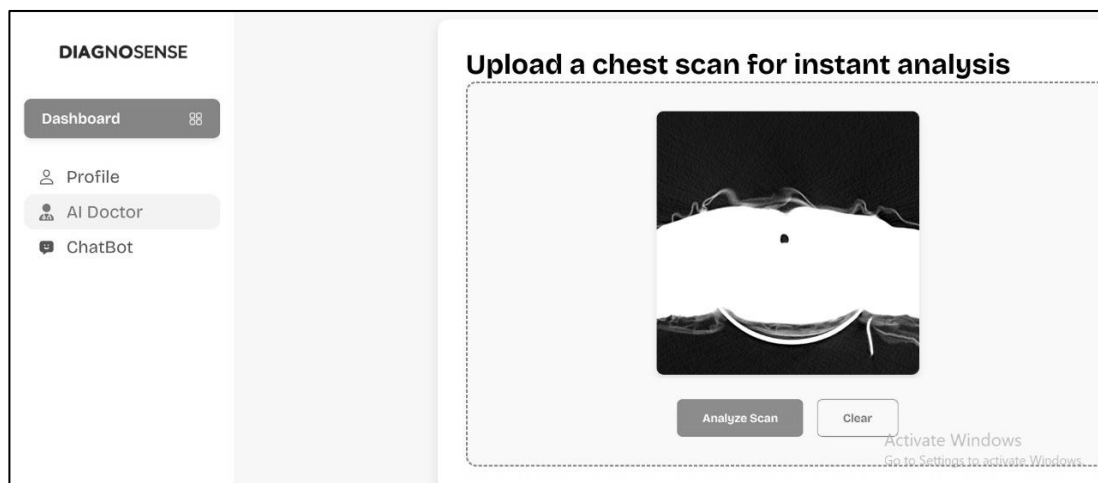
Gender Role
Male PATIENT

Activate Windows
Go to Settings to activate Windows.

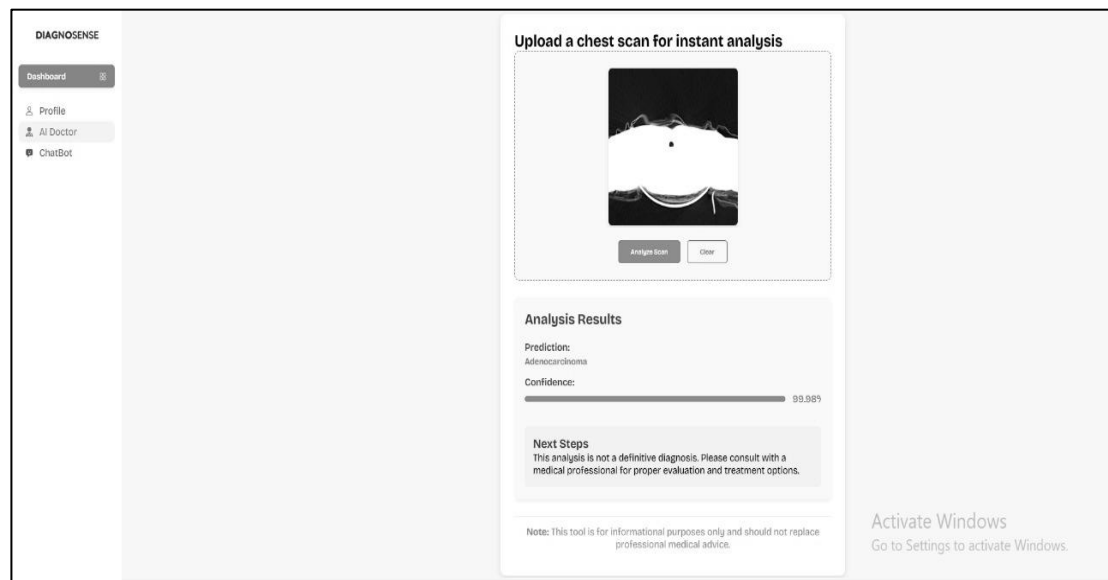
- By clicking AI Doctor, Diagnostic tool will be loaded and user can see a button prompting to upload file. It support multiple formats but preview is only available for JPG and PNG because of size.



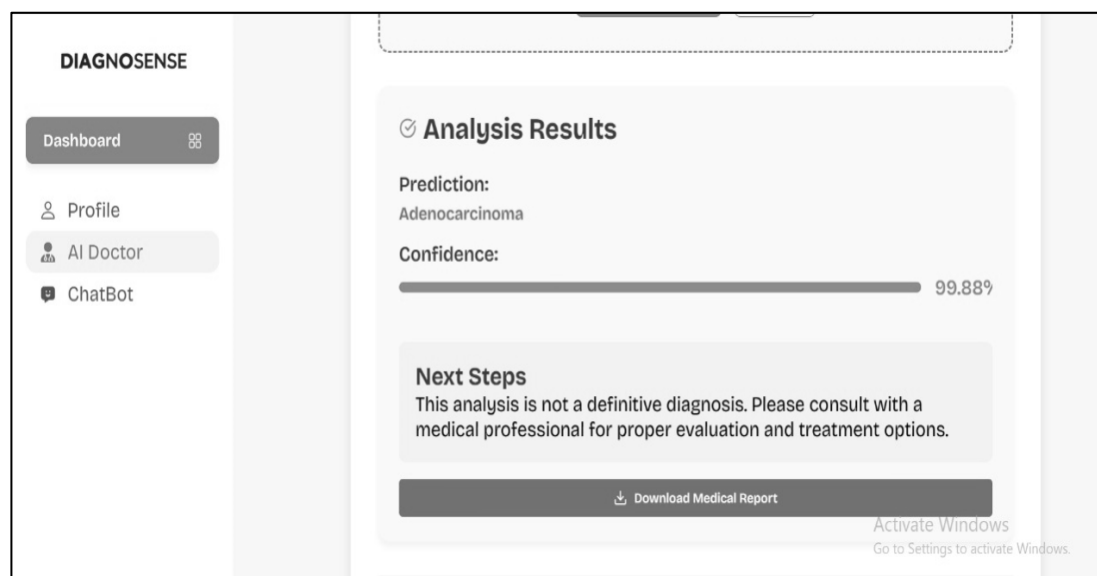
- Upon uploading, preview will show uploaded image following buttons. By clicking on analyse scan, image is send for analysis to trained model.



- After the prediction of model, predicted result class along with confidence will be displayed.



- Once the diagnosis is completed and results are displayed, user can also download predicted output as a PDF report for offline use.



- Report will contain all necessary data of user, entered while creating profile along with results of diagnosis.

DIAGNOSENSE
 AI-Powered Medical Analysis

Medical Scan Analysis Report

Generated on: 5/23/2025, 3:49:20 PM

Patient Information

Patient ID: 4220123456789
 Name: David Doe
 Date of Birth: 1990-05-15T00:00:00.000Z
 Referring Physician: AI Doctor

Analysis Results


Scan Type: Chest X-Ray
 File Name: 1-01.png
 File Size: 92.78 KB

AI Prediction: Adenocarcinoma
Confidence: 99.98%

- For learning more about predicted class or to explore any query related to lung, user can access chatbot by clicking it from side bar.

DIAGNOSENSE

AI Assistant



How can I help you today?

Type your question here...

Activate Windows
 Go to Settings to activate Windows.

CHAPTER 5

RESULTS AND DISCUSSIONS

5.1 Evaluation Results

To effectively evaluate the performance of proposed hybrid ResNet50V2-ViTAttention trained model, both quantitative measures and visualization evaluation were used. Both assessments were performed on a separate test set, which was not encountered during training and validation. The results are structured into three major themes: quantitative metrics, visual training dynamics, and interpretability analysis.

5.1.1 Quantitative Metrics

The assessment of the suggested model is based on traditional machine learning measures, such as the confusion matrix, classification report, and ROC-AUC evaluation. The confusion matrix (Figure 5.1) plots the comprehensive classification results. The model perfectly labeled 236/236 A001-M and 228/228 B000-M cases (no mislabels), with E001-M having 229 correct and 8 incorrect as A001-M (and 1 as G025-M), and G025-M having 210 correct with a few mislabels (3 as A001-M, 3 as B000-M, 5 as E001-M). These findings suggest class A001-M and B000-M were perfectly separated, while E001-M and G025-M had insignificant confusion (may be due to similar patterns). The minimal false negatives and positives match the high per-class precision and recall (≥ 0.96). In summary, the confusion matrix confirms that the hybrid model performs near-perfect discrimination for this task.

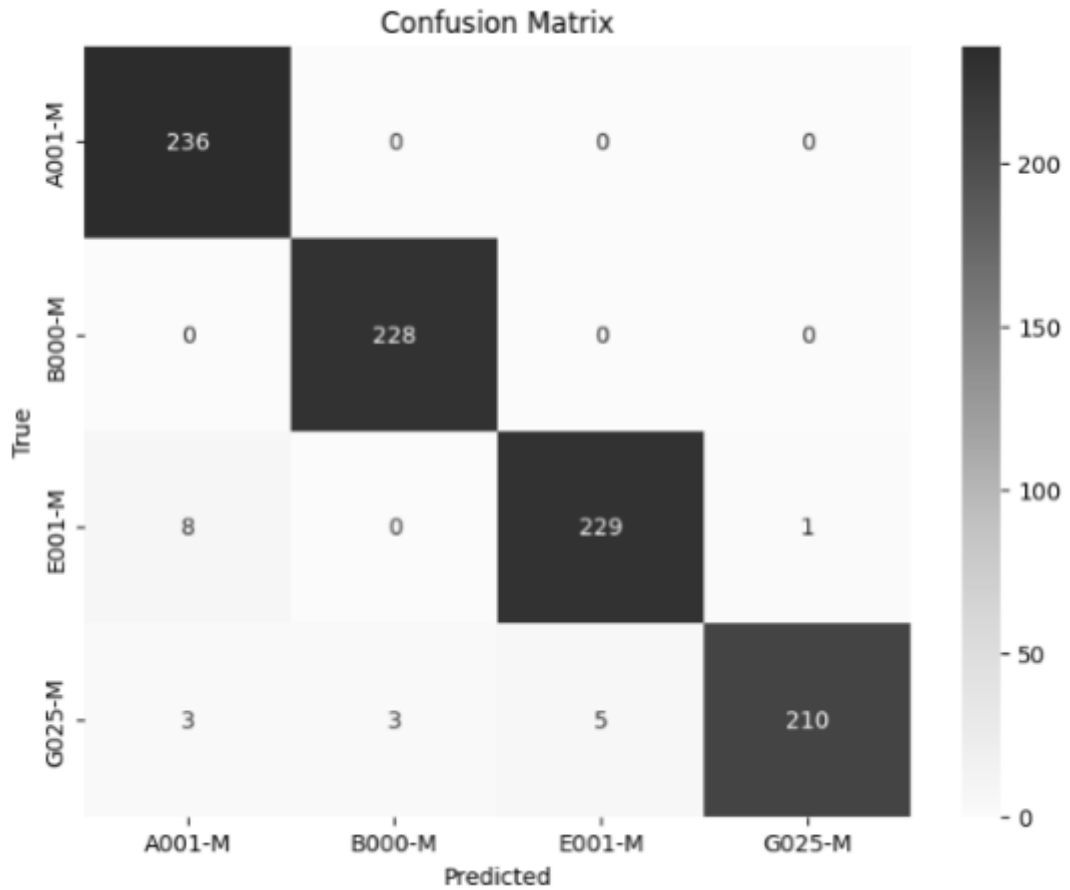


Figure 5.1: Confusion Matrix

Confusion matrix of four-class lung disease classification on the test set. The diagonal values represent correct predictions per class (e.g. 236 for A001-M), and off-diagonals represent misclassifications (e.g. 8 E001-M cases classified as A001-M, etc.). The dominance near the diagonals reflects that most samples are properly labeled, with very little error. The high true positive value and low misclassification value support the high precision and recall values.

Overall accuracy reached **98%** on the test set, with consistently high precision, recall, and F1-scores across all classes. Precision, recall, and F1-score are defined as:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \quad \text{F1} = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where TP , FP , FN denote true positives, false positives, and false negatives, respectively. Using these definitions, our model achieves following results.

Table 5.1: Classification Report

	Precision	Recall	F1-Score	Support
<i>A001-M</i>	0.96	1.00	0.98	236
<i>B001-M</i>	0.99	1.00	0.99	228
<i>E001-M</i>	0.98	0.96	0.97	238
<i>G025-M</i>	1.00	0.95	0.97	221
Accuracy			0.98	923
Macro Avg	0.98	0.98	0.98	923
Weighted avg	0.98	0.98	0.98	923

Macro-avg and weighted-avg F1-scores indicate balanced class performance. Tabular values prove that the model has extremely low false positive as well as false negative rates in all categories, which validates secure detection of signals of lung diseases.

5.1.2 Visualizations

Training and validation accuracy (left) and loss (right) at epochs while the model learns in the feature-extraction stage. The model attains $\approx 90\%$ training accuracy (blue) and $\approx 82\%$ validation accuracy (green) at epoch 6. Both training and validation loss decrease markedly, but signs of overfitting can be seen.

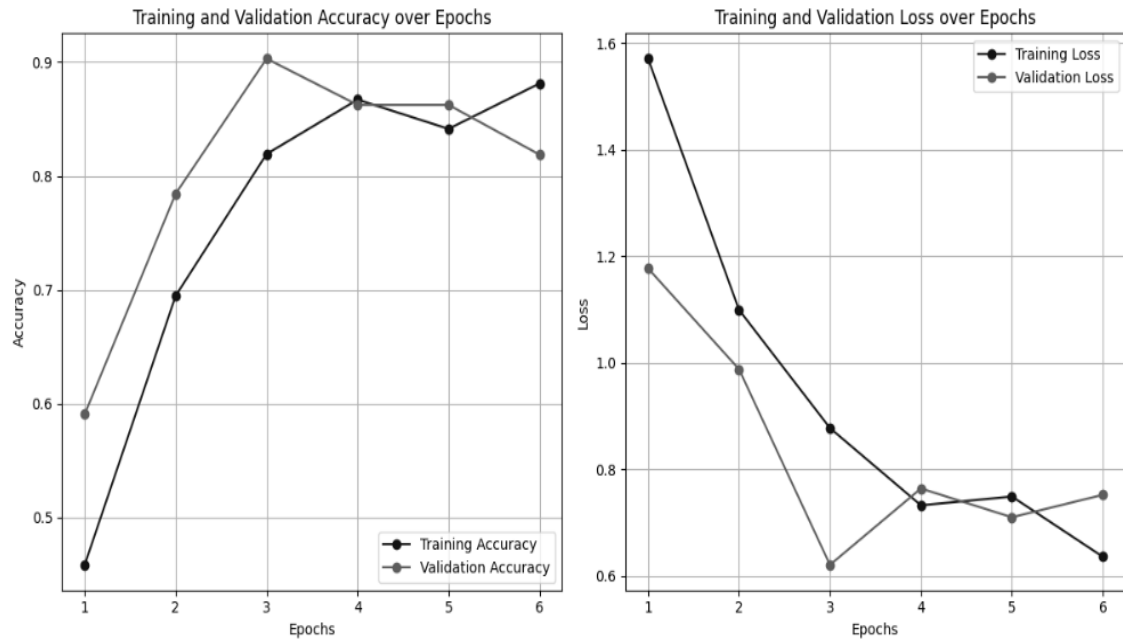


Figure 5.2: Training and Validation Graph

Figure 5.2 is the training curves for the feature-extraction phase (first stage) when only the added classifier and attention layers were trained (ResNet backbone frozen). Training accuracy (blue) increases rapidly to ~ 0.88 , with validation accuracy (green) reaching a peak of ~ 0.90 at epoch 3 before oscillating slightly to ~ 0.82 at epoch 6. Loss curves (right) for training (blue) and validation (green) fall from ~ 1.6 to ~ 0.65 . The large decrease in both training and validation loss suggests good feature learning. The difference between training and validation accuracy is small, suggesting overfitting. These plots show that the early training phase set a good baseline performance but model requires more tuning to generalize more.

Figure 5.3 indicates following the fine-tuning process, wherein unfreezing and retraining the subsequent layers of the pre-trained ResNet50V2 model enables adapting learned features to the particular characteristics of the lung cancer dataset. This adaptation enables the model to learn more general and relevant features, resulting in better predictive capabilities. Conversely, by selectively tweaking the weights of the pre-trained model to adapt more specifically to the particularity of the lung cancer images, fine-tuning adapts the model to the task at hand, producing a more accurate and consistent diagnostic instrument.

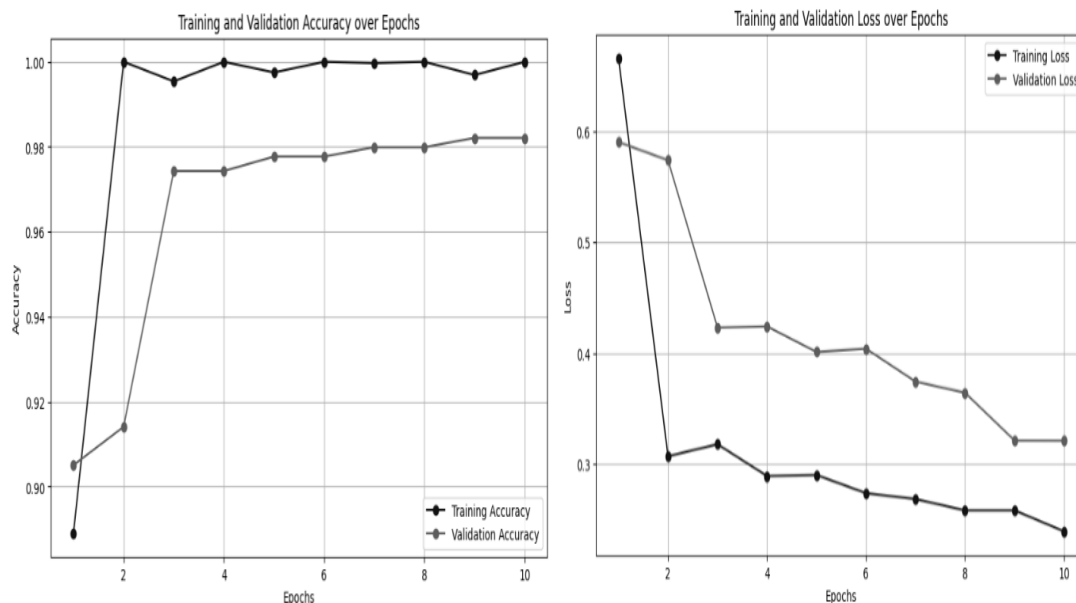


Figure 5.3: Training and Validation Graph (Fine-Tuned)

The accuracy plot (left) after fine-tuning shows a tremendous improvement in the model's classification accuracy. The training accuracy (blue line) shoots up and settles at around 1.00, showing almost perfect classification of the training set for all epochs. At the same time, the validation accuracy (green line) also shows a large improvement, settling at around 0.98, which indicates the model's strong capacity to generalize to new data. The loss plot (right) after fine-tuning shows the model's improved optimization and stability. The training loss (blue line) drops sharply and converges to a very low point, approximately 0.2, indicating that the model is highly accurate in classifying training data. At the same time, the validation loss (green line) also drops substantially and converges to a low value, approximately 0.32, reflecting good generalization and a good match between the model's performance on training and unseen data. The low difference between the training and validation accuracy and loss curves highly indicates that overfitting has been successfully addressed, with the model being in a well-balanced condition of high accuracy and low error on both training and validation sets. This trend reflects the effectiveness of fine-tuning in optimizing the model's parameters for best performance and generalization.

To further evaluate the discriminative ability of the model, the Receiver Operating Characteristic (ROC) curve was drawn and is depicted in **Figure 5.4**. The curve rises sharply towards the top-left quadrant of the plot, revealing that the model retains a high rate of true positives even at different threshold levels.

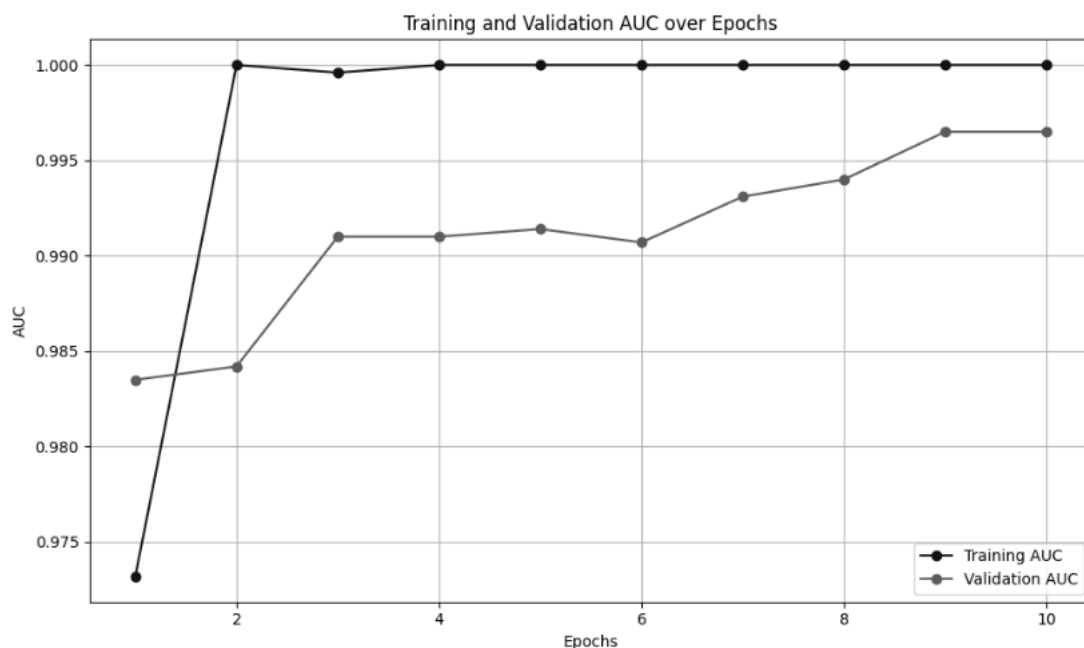


Figure 5.4: Training and Validation AUC Graph

The Area Under the Curve (AUC) is close to 1.0, validating the model's high ability to discriminate between non-cancerous and cancerous cases. The ROC-AUC curve is a threshold-independent performance measure, which is especially useful while tuning the model for practical application in different clinical settings.

Training and validation metrics across epochs (Figure 5.5) in fine-tuning. Training AUC: solid blue line, validation AUC: dashed blue; red/green curves: precision and recall (solid: training, dashed: validation). By epoch 2–3, all the metrics have converged to ≈ 0.98 –1.00 and plateau, showing convergence and extremely consistent generalization.

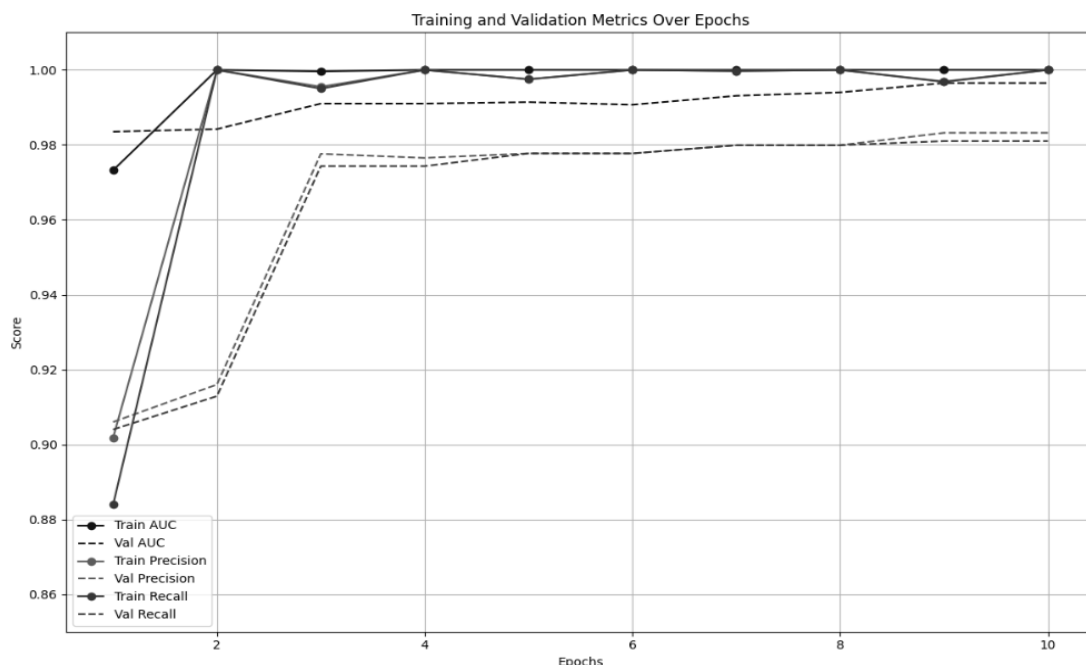


Figure 5.5: Training and Validation Metrics Graph

In fine-tuning stage, all network layers (or the final layers of ResNet-50) were unfrozen and trained. **Figure 5.5** shows AUC (area under ROC curve) for training (solid blue) and validation (dashed blue) over epochs. Follows a steep increase, and training AUC converges to 1.00 as early as epoch 2 and stays there, whereas validation AUC rises to ≈ 0.99 by epoch 10. Also plots precision and recall for training (solid red/green) and validation (dashed red/green). By epoch 3, all these values settle around 0.97–1.00 for both. The high rate of convergence and the approximate coincidence of the training and validation curves suggest good generalizability of the model; the extremely high AUC supports very high class separability. Both, combined, these learning curves indicate that fine-tuning provides low overfitting and maximizes performance.

5.1.3 Interpretability

Contemporary clinical machine learning algorithms should not only work well but also provide interpretability, so clinicians can comprehend the rationale behind a model's choice. To satisfy this requirement, we utilized Local Interpretable Model-Agnostic Explanations (LIME), a post-hoc interpretability method that explains any classifier's prediction in an interpretable and faithful way. LIME explanation example for a CT image. Yellow contours point to areas labeled by the model as significant for classification. The areas highlighted represent abnormal lung areas, showing the model is centered on clinically relevant features in the PET/CT image.

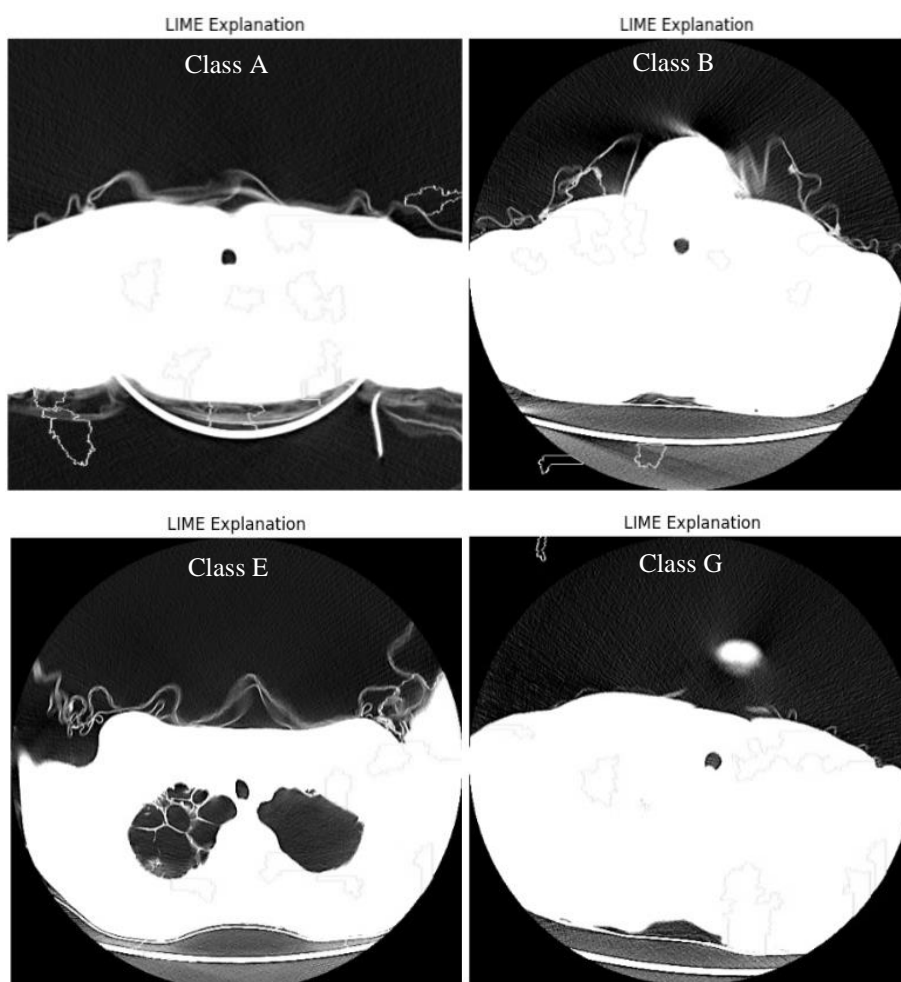


Figure 5.6: LIME Interpretation

Figure 5.6 is a sample LIME (Local Interpretable Model-agnostic Explanations) visualization of a CT slice. The yellow borders indicate super pixel areas that had the greatest impact on the model's prediction. We can see that these highlighted areas are where there is abnormal tissue density in the lungs. This suggests that the model's predictions are based on relevant pathology cues and not on irrelevant features. These explainable outputs confirm that the hybrid model is paying attention to proper regions of interest. In real-world settings, this interpretability can build clinician confidence, as the model's attention maps correspond to human-understood markers of disease. This localized visual explanation is especially important in high-stakes uses such as cancer diagnosis. Transparency is assured, trust is built with clinical practitioners, and there is a method for model validation and error auditing. The LIME-based explanation not only ensures the model is making well-informed decisions but also facilitates cross-verification of predictions with radiologist judgment.

The model reflects excellent classification performance, underpinned by solid training behavior, stable metrics, and easy visualization tools. Training curves show productive optimization and very little overfitting. Quantitative measures validate balanced and exact classification. More importantly, incorporation of ROC-AUC and LIME explanations augments the clinical usefulness of the model by providing performance transparency as well as interpretability of decisions. These facets combined support the feasibility of the model for application in real-life scenarios in the field of automatic lung cancer diagnosis.

5.1.4 Comparison With State-of-the-Art Methods

The proposed hybrid CNN–ViT model invariably produced state-of-the-art classification performance. Overall accuracy at 98% and well-balanced F1-scores (~ 0.98 macro-average) confirm that including both local and global features worked effectively. Against recent work, our method possesses competitive or even better performance. For instance, Hammad et al.[39] obtained 93.1% accuracy on a CT classification task of a lung cancer subtype using a Grad-CAM explanation and a custom CNN [38]; our model performs better than this, probably because it utilizes both PET and CT modalities and attention mechanisms. Likewise, a transformer-based

approach on Lung-PET-CT-DX obtained tumor staging accuracies $\approx 97\%$ [40], which is similar to our classification accuracy. The earlier hybrid CNN–ViT for chest X-rays had $\sim 98\text{--}99\%$ accuracy for binary classification of lung diseases [41]; our multi-class outcomes ($\sim 98\%$) are comparable, highlighting the effectiveness of hybrid configurations (or higher, considering multi-class difficulty).

The studies considered here mostly employed CNN-based models like VGG16, MobileNetV2, and EfficientNet, commonly coupled with shallow learners like RF and XGBoost. Gupta et al. [32] used transfer learning and attained 89.7% accuracy but did not attempt attention mechanisms or Transformer-based models. Singh et al. [25] implemented a tailored CNN trained from scratch with reduced accuracy (86.5%) and increased overfitting owing to less augmentation. Kumar et al. (21) integrated EfficientNet with a Random Forest classifier and achieved 88.2% accuracy but had difficulty with early-stage detection as a result of low recall. Ahuja et al. [42] emphasized sensitivity, employing MobileNetV2 to achieve 87% sensitivity, although the model was not robust in detecting metabolic variations. Iqbal et al. [40] reported the highest accuracy (90.1%) through a CNN ensemble but, as other CNN-only models, did not include mechanisms for modeling long-range dependencies and multi-scale attention essential for complicated diseases such as cancer or fibrosis.

Hybrid architecture imparts major benefits. ResNet50V2 CNN layers well capture high-resolution local patterns (edges, textures) in every slice, and the ViT-based attention block incorporates global context patterns from the whole image. As proved in earlier literature, vanilla ViTs are indeed good at modeling long-range dependencies, but would normally need to have access to very large dataset. Our model circumvents this issue by marrying ViT's self-attention with a pre-trained ResNet backbone. This synergy is seen in the outcome: the two pathways complement one another, producing richer features than either in isolation [41]. This is made clear by the confusion matrix: classes with fine interclass distinctions (E001-M vs. G025-M) are correctly resolved largely, indicating global context assists in discriminating them. Two-stage training (feature extraction followed by fine-tuning) also worked well. Freezing the CNN first enabled the attention block to settle, and subsequent unfreezing resulted in steady improvement without overfitting (see Figures 5.2,5.3,5.5). This

curriculum-type training is recognized to enhance convergence in transfer learning configurations, particularly for small medical data.

Data augmentation and visualization of attention also enhance robustness. We used rotations, shifts, zooms and flips in training, increasing the effective dataset as suggested in surveys [29]. This probably helped the model generalize (as evidenced by high validation scores). The LIME heatmap (Figure 5.6) also supports that the model learned medically-meaningful patterns: the marked lung lesions coincide with anticipated disease areas. Together with the high F1-scores, this indicates that the network is not over-relying on spurious artifacts. Furthermore, the LIME and attention-based interpretability methods guarantee that predictions are transparent, which is essential for clinical uptake.

Clinically, the hybrid model's explainability and performance are encouraging. On the Lung-PET-CT-DX dataset, our model has near-perfect disease label detection. In practice, this would mean better computer-aided diagnostic (CADx) systems. High recall (sensitivity) is crucial to prevent missed cancers, and our model's recall ~ 0.98 implies very few cases are missed. The high precision reduces false positives, meaning fewer unnecessary follow-ups. The attention maps of the model can be shown to radiologists in conjunction with predictions, possibly directing their attention and establishing trust. Additionally, since ResNet50V2 is quite efficient and the ViT block is light, the system can be implemented without too much computation, a boon in clinical environments.

Table 5.2: Comparison of Proposed Method with state-of-the-art Methods

Article	Model Used	Evaluation Metrics	Limitations
Gupta et al. (2023) [32]	VGG16, ResNet50, XGBoost	Accuracy, AUC, Prec Best: 89.7% Acc.	No ViTs, no hybrid learning
Iqbal et al. (2023) [40]	Ensemble CNN	Accuracy, F1, ROC-AUC Best: 90.1% Acc.	CNN-only, lacks contextual modeling
Kumar et al. (2024) [25]	EfficientNet, RF	Accuracy, F1-score Best: 88.2% Acc., 0.87 F1	Low recall for early-stage cancer
Ahuja et al. (2024) [42]	MobileNetV2	Precision, Sensitivity Best: 87% Sens.	Inadequate for PET feature learning
Siyuan Tang et al. (2025) [43]	Improved SwinUNet with CNN-Transformer dual encoder, CBAM, DAM, SDIM, semi-supervised learning	ACC, Precision, Recall, DSC, IoU, 95HD, ASD Best: ACC: 96.28% DSC: 93.72% IoU: 89.02% 95HD: 3.42mm ASD: 1.55mm	Relatively weak segmentation on small cell carcinoma subset, Requires accurate pseudo-labeling
Fatih Aksu, Fabrizia Gelardi, Arturo Chiti, Paolo Soda. (2025) [44]	Custom 3D multimodal CNN with multi-stage intermediate fusion blocks based on ResNet architecture	Accuracy, AUC, Gmean used for class imbalance handling Best: ACC: 0.724 AUC: 0.681 Gmean: Not specified explicitly,	Moderate classification performance, Computationally intensive, Slight spatial misalignment issues
Proposed Model	Hybrid CNN–Transformer architecture: ResNet50V2 (pretrained) Custom ViTAttention Block	Accuracy, Precision, Recall, F1-score, AUC Best: Accuracy: 98% Macro F1-score: 0.98 AUC: 0.99+	Uses 2D slices, ViT block not pretrained, Limited interpretability (post hoc only), No clinical workflow validation

In short, by combining the established power of CNN feature extractors with the global reasoning of Transformers, our hybrid model achieves outstanding diagnostic accuracy with explainable outputs. It is comparable to current methods in the literature, yet it only needs moderate data due to transfer learning and augmentation. We believe that this method has tremendous potential for real-world lung cancer detection pipelines, where fast and trustworthy interpretation of PET/CT scans is crucial for patient outcomes.

5.1.5 Identified Limitations

Although the presented hybrid ResNet50V2–ViTAttention model performed well in high classification accuracy and robust generalizability on the LUNG-PET-CT-DX dataset, multiple limitations must be taken into account. Firstly, the dataset is derived from a single institution and may restrict generalizability to heterogeneous clinical settings. External validation using multi-center datasets should be performed in future work to measure robustness across disparate imaging protocols and population demographics.

Second, while the model makes use of both PET and CT modalities, it receives 2D image slices as input instead of 3D volumetric data. This potentially causes loss of valuable spatial continuity associated with medical imaging. Using 3D CNNs or volumetric transformer-based models may be able to capture such spatial relationships better.

Furthermore, the Vision Transformer block was trained in a scratch mode with no access to large-scale pretraining, which can restrict its capacity for modeling intricate global dependencies. Adding self-supervised or pre-trained transformer backbones would improve performance. The model also has no specific mechanisms for handling class imbalance; future applications can be improved using weighted loss functions or resampling techniques. While LIME was used for interpretability, post hoc methods like it might not capture the causal reasoning behind predictions. More inherent interpretability methods like attention-based saliency or clinically-guided attention mechanisms might enhance clinical trust.

CHAPTER 6

CONCLUSION AND RECOMMENDATIONS

6.1 Conclusion

This project effectively built and implemented a hybrid deep learning framework for lung disease classification and anatomical support, addressing both diagnostic precision and user interpretability. By combining the spatial understanding of ResNet50V2 with the global contextual modelling capability of a Vision Transformer (ViT Attention), the system was capable of delivering robust performance in classifying PET/CT images over a variety of lung disease classes. The use of a two-stage training approach further enhanced the model's stability and generalization on intricate medical imaging data.

One of the major contributions of this study is the integration of a domain-specific chatbot to answer user queries regarding lung anatomy. Built using a Retrieval-Augmented Generation (RAG) approach, the chatbot leverages in-grained medical knowledge and large language model capability to give accurate, grounded responses in real-time. This integration transforms the platform from being an exclusive diagnostic tool to an interactive resource that can assist clinical education and decision making. The entire solution was made accessible through a safe, intuitive web interface, and clinicians, radiologists, and researchers could intuitively interact with the system. Certain key features such as real-time inference, explain ability through LIME, and privacy-sensitive design positioned the system for application in medical and research environments in real-world conditions.

In short, the project demonstrates that the fusion of leading-edge deep learning methods and interactive AI can produce smarter, comprehensible, and scalable medical tools. Besides enabling early and precise diagnosis of lung disease, the system also inspires deeper understanding of lung anatomy to eventually lead to better-informed clinical decisions and better patient care.

6.2 Future Work

Although the developed system possesses great potential for accurate lung disease classification and interactive anatomical guidance, there are several areas for enhancement that can be used to further enhance its clinical usefulness, scalability, and flexibility.

- **3D Volumetric Analysis:** Subsequent versions of the model may make use of full 3D volumetric PET/CT data instead of 2D slices. This would allow the system to improve spatial relationship registration between several layers of scans, improve diagnostic accuracy, and identify deeper or multi-focal lesions.
- **Multi-Modality Fusion:** Enhancing the system to handle PET and CT inputs separately in independent streams before merging their representations may produce a more precise and inclusive classification, especially for complex or ambiguous cases.
- **Pretraining the Vision Transformer:** The current Vision Transformer (ViTAttention) block was trained from scratch due to domain specificity. Performance can be improved by pretraining the transformer module on large-scale medical imaging datasets or self-supervised learning techniques before fine-tuning on the lung-specific dataset.
- **Expansion of Chatbot Scope:** Though currently focused on lung anatomy, its knowledge base could be expanded for other organ systems or disease-based data. Further, the provision of context-sensitive follow-up processing and

visualized explanation capabilities (e.g., pointing to illustrations of anatomy) would also contribute to enhanced user interaction.

- **Interoperability with PACS and DICOM Standards:** To enable direct use in clinical settings, future releases can be provided to interface with hospital Picture Archiving and Communication Systems (PACS) and provide full DICOM standard support so that access to real-time clinical imaging data can be provided seamlessly.
- **Mobile Compatibility and Offline Mode:** Development of a mobile-optimized version or hybrid app would allow the system to be accessed on tablets or phones by clinicians at the point of care. Offline functionality, especially for the chatbot, would allow use in low-resource settings or field operations.
- **Clinical Trials and Real-World Validation:** Finally, in order to reach clinical trust and regulatory readiness, the system has to be proven in real clinical environments through pilot studies or trials. This would allow for performance benchmarking, human-AI collaboration assessment, and feedback collection for continuous improvement.

REFERENCES

- [1] Bray F, et al., "Global cancer statistics 2018: GLOBOCAN estimates," *CA: A Cancer Journal for Clinicians*, 2018.
- [2] Sung, H., et al. "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries." *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, 2021, pp. 209–249.
- [3] American Lung Association, "Lung Cancer Survival Rates," 2020.
- [4] Quekel LG, et al., "Missed lung cancer on chest radiographs," *Chest*, 1999.
- [5] Brady A, "Error and discrepancy in radiology," *Insights into Imaging*, 2012.
- [6] Berlin L, "Radiologic errors and malpractice: a blurry distinction," *American Journal of Roentgenology*, 2007.
- [7] Reiner BI, et al., "Radiologist workload: update," *Journal of the American College of Radiology*, 2010.
- [8] American Cancer Society, *Cancer Facts & Figures 2023*. [Online]. Available: <https://www.cancer.org/>
- [9] A. Jemal et al., "Global cancer statistics," *CA: A Cancer Journal for Clinicians*, vol. 70, no. 1, pp. 7–30, 2020.
- [10] D. Aberle et al., "Reduced lung-cancer mortality with low-dose computed tomographic screening," *New England Journal of Medicine*, vol. 365, no. 5, pp. 395–409, 2011.
- [11] R. Sone et al., "Missed lung cancers on chest radiographs: Results of a computer-aided detection system," *Radiology*, vol. 241, no. 2, pp. 568–575, 2006.
- [12] H. J. MacMahon et al., "Guidelines for management of small pulmonary nodules detected on CT scans: A statement from the Fleischner Society," *Radiology*, vol. 237, no. 2, pp. 395–400, 2005.
- [13] Clark K, et al., "LUNG-PET-CT-DX Dataset," The Cancer Imaging Archive (TCIA), 2020.

- [14] S. Makaju, P. W. C. Prasad, A. Alsadoon, A. K. Singh, and A. Elchouemi, "Lung Cancer Detection using CT Scan Images," *Procedia Computer Science*, vol. 125, pp. 107–114, Jan. 2018, doi: 10.1016/j.procs.2017.12.016.
- [15] C. Zhang et al., "Toward an expert level of lung cancer detection and classification using a deep convolutional neural network," *The Oncologist*, vol. 24, no. 9, pp. 1159–1165, Apr. 2019, doi: 10.1634/theoncologist.2018-0908.
- [16] A. Elnakib, H. M. Amer, and F. E. Z. Abou-Chadi, "Early Lung Cancer Detection using Deep Learning Optimization," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 16, no. 06, pp. 82–94, May 2020, doi: 10.3991/ijoe.v16i06.13657.
- [17] M. S. Al-Huseiny and A. S. Sajit, "Transfer learning with GoogLeNet for detection of lung cancer," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 2, p. 1078, May 2021, doi: 10.11591/ijeecs.v22.i2.pp1078-1086.
- [18] N. Nasrullah, J. Sang, M. S. Alam, M. Mateen, B. Cai, and H. Hu, "Automated Lung Nodule Detection and Classification Using Deep Learning Combined with Multiple Strategies," *Sensors*, vol. 19, no. 17, p. 3722, Aug. 2019, doi: 10.3390/s19173722.
- [19] S. Tang et al., "Segmentation of PET/CT Lung Cancer Lesion Images via Semisupervised Improved SwinUNet Model," *Ssrn*, Jan. 2025, doi: 10.2139/ssrn.5069919.
- [20] F. Aksu, F. Gelardi, A. Chiti, and P. Soda, "Multi-stage intermediate fusion for multimodal learning to classify non-small cell lung cancer subtypes from CT and PET," *Pattern Recognition Letters*, Apr. 2025, doi: 10.1016/j.patrec.2025.04.001.
- [21] K.-Y. Huang, C.-L. Chung, and J.-L. Xu, "Deep learning object detection-based early detection of lung cancer," *Frontiers in Medicine*, vol. 12, Apr. 2025, doi: 10.3389/fmed.2025.1567119.
- [22] Setio, A.; Ciompi, F.; Litjens, G.; Gerke, P.; Jacobs, C.; Riel, S.; Wille, M.M.; Naqibullah, M.; Sanchez, C.I.; van Ginneken, B. Pulmonary nodule detection in CT images: False positive reduction using multi-view convolutional networks. *IEEE Trans. Med. Imaging* 2016, 35, 1160–1169.
- [23] Chen, C.; Zhou, K.; Zha, M.; Qu, X.; Xiao, R. An effective deep neural network for lung lesions segmentation from COVID-19 CT images. *IEEE Trans. Ind. Inform.* 2021, 17, 6528–6538.
- [24] A. R. W. Sait, "Lung cancer detection model using deep learning technique," *Applied Sciences*, vol. 13, no. 22, p. 12510, Nov. 2023, doi: 10.3390/app132212510.

- [25] Wang, D., et al. (2017). Zoom-in-Net: Deep Mining Lesions for Diabetic Retinopathy Detection. MICCAI.
- [26] Rajpurkar P, et al., "CheXNet: Radiologist-Level Pneumonia Detection," arXiv, 2017.
- [27] M. M. Ansari et al., "SVMVGGNET-16: A novel machine and deep learning based approaches for lung cancer detection using combined SVM and VGGNET-16," Current Medical Imaging Formerly Current Medical Imaging Reviews, vol. 21, Jan. 2025, doi: 10.2174/0115734056348824241224100809.
- [28] Shorten C, Khoshgoftaar TM, "A Survey on Image Data Augmentation for Deep Learning," Journal of Big Data, 2019.
- [29] Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data.
- [30] Taylor, L., & Nitschke, G. (2018). Improving Deep Learning with Generic Data Augmentation. arXiv preprint arXiv:1708.06020.
- [31] Jaderberg, M., Simonyan, K., & Zisserman, A. (2015). Spatial Transformer Networks. Advances in Neural Information Processing Systems (NeurIPS).
- [32] L. Wang, C. Zhang, Y. Zhang, and J. Li, "An automated diagnosis method for lung cancer target detection and subtype Classification-Based CT scans," Bioengineering, vol. 11, no. 8, p. 767, Jul. 2024, doi: 10.3390/bioengineering11080767.
- [33] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. NeurIPS.
- [34] Lim, S., et al. (2019). Fast AutoAugment. NeurIPS.
- [35] Goodfellow, I., et al. (2016). Deep Learning. MIT Press (Chapter on Regularization).
- [36] I. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts," in Proc. Int. Conf. Learning Representations (ICLR), 2017. [Online]. Available: <https://arxiv.org/abs/1608.03983>
- [37] J. Brownlee, How to Stop Training Deep Neural Networks at the Right Time Using Early Stopping, Machine Learning Mastery, 2018. [Online]. Available: <https://machinelearningmastery.com/how-to-stop-training-deep-neural-networks-at-the-right-time-using-early-stopping/>
- [38] K.-Y. Huang, C.-L. Chung, and J.-L. Xu, "Deep learning object detection-based early detection of lung cancer," Frontiers in Medicine, vol. 12, p. 1567119, 2025.

- [39] M. Hammad, M. ElAffendi, A. a A. El-Latif, A. A. Ateya, G. Ali, and P. Plawiak, "Explainable AI for lung cancer detection via a custom CNN on CT images," *Scientific Reports*, vol. 15, no. 1, Apr. 2025, doi: 10.1038/s41598-025-97645-5.
- [40] Iqbal MS, Heyat BBM, Parveen S, et al. Progress and trends in neurological disorders research based on deep learning. *Comput Med Imaging Graph* 2024; 116: 102400. [<http://dx.doi.org/10.1016/j.compmedimag.2024.102400>] 38851079]
- [41] Y. Hadhoud et al., "From Binary to Multi-Class Classification: A Two-Step hybrid CNN-VIT model for chest disease classification based on X-Ray images," *Diagnostics*, vol. 14, no. 23, iqbalp. 2754, Dec. 2024, doi: 10.3390/diagnostics14232754.
- [42] A. Wehbe, S. Dellepiane, and I. Minetti, "Enhanced lung cancer detection and TNM staging using YOLOv8 and TNMClassifier: An integrated deep learning approach for CT imaging," *IEEE Access*, 2024.
- [43] Y. Miao, N. Wang, L. Liu, Y. Qu, G. Yu, Q. Ji, Q. Bao, and J. Zhao, "Segmentation of PET/CT lung cancer lesion images via semisupervised improved SwinUNet model,"
- [44] F. Aksu, F. Gelardi, A. Chiti, and P. Soda, "Multi-stage intermediate fusion for multimodal learning to classify non-small cell lung cancer subtypes from CT and PET," *arXiv preprint arXiv:2501.12425*, 2025.