Available online at sjuoz.uoz.edu.krd

Science Journal of University of Zakho

Vol. 13, No.4 pp599–618 October-December,2025

p-ISSN: 2663-628X
e-ISSN: 2663-6298

A PRUNED VGG16 WITH HYBRID PREPROCESSING AND DATA BALANCING FOR ROBUST AND INTERPRETABLE LUNG CANCER CLASSIFICATION

Marwa Salih Ramadhan[1*], and Mohammed Ahmed Shakir[2]

[1] Technical Administrative College, Duhok Polytechnic University, Kurdistan Region, Iraq
[1, 2] Electrical and Computer Engineering Department, College of Engineering, University of Duhok, Kurdistan Region, Iraq

[*]Corresponding author email: marwaeng.abdo@gmail.com

Received: 29 May 2025 Accepted:14 Jul 2025 Published:8 Oct 2025 https://doi.org/10.25271/sjuoz.2025.13.4.1597

ABSTRACT:

Lung cancer is the most common and deadliest type of cancer globally, creating a critical need for diagnostic tools that are not only accurate but also practical for clinical integration. This study introduces a robust, computationally efficient, and interpretable deep learning framework using Computed Tomography (CT) images to address limitations in existing models, such as high computational costs, poor data quality, and a lack of transparency. Our approach utilizes a VGG16 architecture, streamlined through structured pruning, which reduced the parameter count from 138.3M to 26.6M without compromising performance. We developed a hybrid pipeline with dual filtering and adaptive CLAHE to enhance image quality, while data diversity and imbalance were mitigated using hybrid augmentation and SMOTE. The model was trained with a rigorous strategy, including four-fold cross-validation and dual-phase fine-tuning with a dynamic learning rate, ensuring stable convergence. On a primary single-source dataset, the model achieved a test accuracy of 0.9910 and a Matthews Correlation Coefficient (MCC) of 0.9845. To validate real-world applicability, the framework was tested on a large multi-source dataset, demonstrating strong generalization with a balanced accuracy of 0.9693 and an MCC of 0.9427. Model interpretability was confirmed using Grad-CAM visualizations to highlight clinically relevant regions. This framework provides a highly accurate, computationally efficient, and generalizable solution with significant potential for clinical deployment as a reliable diagnostic aid.

KEYWORDS: Lung cancer; Deep Learning (DL); Computed Tomography (CT); Transfer learning; Hybrid Preprocessing; VGG16, Synthetic Minority Oversampling Technique (SMOTE); Grad-CAM.

1. INTRODUCTION

Background:

Over the past decades, lung cancer has increased dramatically and has remained the most prevalent and deadly malignancy worldwide (Bray *et al.*, 2024). According to the International Agency for Research on Cancer, in 2022, lung cancer accounted for 12.4% of all new cancer cases (2.5 million) and 18.7% of cancer deaths (1.8 million). Projections estimate a 77% increase in these statistics by 2050 (Leiter *et al.*, 2023). The diagnosis begins when pulmonary nodules are detected on medical imaging; these nodules indicate abnormal and uncontrolled growth of lung cells and are classified as either benign (slow-growing, non-metastatic) or malignant (rapidly growing and metastatic). This kind of cancer is often detected at advanced stages when treatment options are limited and survival chances are lower (WHO, 2023). In this regard, it would seem that improving early detection methods could reduce the high death rate due to lung cancer.

It is worth noting that several methods exist for detecting lung nodules in the chest. Among these, Computed Tomography (CT) is considered the most effective because it usually produces high-resolution, cross-sectional images of the lungs using X-rays. For this reason, the provided images offer better visibility of lung nodules compared to standard chest X-rays (Murad *et al*., 2023). Furthermore, a special type of CT, known as Low-Dose Computed Tomography (LDCT), uses significantly less radiation. Despite the lower dose, it still provides accurate detection. Studies have shown that LDCT can reduce lung cancer mortality by around 20% Therefore, it was recommended that LDCT be used as a standard screening tool for high-risk individuals (National Lung Screening Trial Research Team, 2011).

However, manually interpreting CT scans remains time-consuming and may lead to inconsistent diagnoses between experts. Consequently, these limitations may cause treatment delays and negatively impact patient outcomes. Therefore, there is a critical need for automated diagnostic systems to support clinical decision-making. Recent advances in deep learning (DL), particularly Convolutional Neural Networks (CNNs), have shown significant promise in medical imaging because they can automatically learn discriminative features directly from raw images. Studies have demonstrated that CNNs can detect and

* Corresponding author

classify lung nodules with accuracy comparable to that of experienced radiologists (Ibrahim & Mahmood, 2023).

Among the many available CNN architectures, VGG16 has become widely adopted for its effectiveness in medical imaging. Its simple and well-organized design includes small 3x3 filters and stacked convolutional layers, allowing it to extract a rich hierarchy of features, from basic edges to complex patterns. Furthermore, its use of ReLU activation adds non-linearity, enhancing the model's overall learning capacity. Ultimately, these extracted features are passed to fully connected layers for classification, making VGG16 a powerful tool for medical analysis (Ardila *et al.,* 2019).

Problem Statement:

Although deep learning (DL) has shown growing success in lung cancer detection and classification, several limitations hinder its clinical deployment. A primary challenge starts from medical datasets, which are often small and imbalanced. Furthermore, each dataset collection contains different variations and noise acquired from different CT scanners, which reduces model reliability and generalization. Another key issue is that standard DL architectures are computationally expensive, making them unsuitable for resource-constrained settings. While lightweight models offer an efficient alternative, they cannot frequently capture the fine-grained patterns required in medical imaging.

Furthermore, the use of weak validation strategies, such as simple train-test splits and evaluation on single-source datasets, often leads to poor generalization on different source data. Additionally, many studies rely overly on accuracy while neglecting more robust metrics, such as F1-score, Matthews Correlation Coefficient (MCC), Cohen's Kappa, and Confidence Intervals (CI), that better reflect true model performance. Moreover, interpretability remains a significant concern. Most DL models function as "black boxes," which can limit clinical trust and acceptance. All these compound challenges underscore the need for DL frameworks that are efficient, interpretable, and rigorously validated for real-world clinical use.

Study Objectives:

This study aims to develop a robust, efficient, and interpretable VGG16 model for multi-class lung cancer classification that generalizes across diverse CT data for reliable clinical use.

Specific Objectives:

1. To mitigate challenges related to data quality and imbalance by implementing an advanced preprocessing pipeline and class-balancing techniques. This will enhance the model's reliability and generalization across diverse and noisy datasets.
2. To reduce the computational cost of the VGG16 architecture through strategic parameter pruning and fine-tuning. Consequently, the resulting model will be efficient enough for deployment in resource-constrained clinical environments.
3. To improve model transparency and build clinical trust by integrating explainability (XAI) methods. These tools will visualize the model's decision-making process, moving beyond a 'black box' approach.
4. To conduct a rigorous and clinically relevant evaluation of the model. Therefore, performance will be assessed using K-Fold cross-validation with comprehensive metrics and validated on multi-source data to ensure its real-world effectiveness.

Study Contributions:

This study makes several key contributions to automated lung cancer detection by addressing common limitations in data preparation, model efficiency, and validation. Our main contributions are:

1. *A Quantitatively Validated Preprocessing Pipeline:* We developed a pipeline that combines hybrid filtering and adaptive contrast enhancement. Its effectiveness was confirmed through comparative analysis and quantitative metrics (PSNR, SSIM) to ensure it improves image quality without introducing artifacts.
2. *A Hybrid Strategy for Data Diversity and Balancing:* We implemented a strategy combining hybrid augmentation with SMOTE oversampling. This creates a more diverse and balanced dataset, which directly improves the model's ability to generalize.
3. *A Lightweight and Efficient Deep Learning Architecture:* We developed a highly efficient model by pruning the pretrained VGG16 architecture and applying a dual-phase fine-tuning strategy. This reduced the model's parameters and memory size by more than 80% without sacrificing accuracy, making it practical for clinical deployment.
4. *A Robust Multi-Layered Validation:* We employed a two-tiered strategy combining K-Fold Cross-Validation and external validation on a large multi-source dataset. This comprehensive approach, using metrics like F1-score, MCC, and 95% CI, demonstrates the model's robustness and real-world generalizability.
5. *Enhanced Clinical Interpretability:* We integrated Grad-CAM to provide visual explanations of the model's predictions. These heatmaps enhance transparency, build clinical trust, and support expert review.

2. LITERATURE REVIEW

To design a clinically robust and generalizable lung cancer classification system, we examined recent deep learning research using CT imaging. This section categorizes key studies by methodology and outlines common limitations—particularly in preprocessing, class balancing, validation, and explainability—that our proposed framework aims to overcome.

Summary of Recent Studies in Lung Cancer Classification:

The Automated lung cancer diagnosis has progressed from traditional machine learning to advanced deep learning pipelines. Current literature highlights diverse architectural approaches, from pretrained CNNs to custom and hybrid models.

Preprocessing and Data Handling:

Effective preprocessing is a critical first step for any diagnostic model. The literature shows a wide range of approaches. Many studies have employed foundational techniques such as image resizing, normalization, and basic filtering (Anand *et al*., 2022; Tandon *et al*., 2022). More advanced studies have incorporated specific denoising filters like Gaussian or median filters (Gupta *et al*., 2023; Ravindra *et al*., 2024) and contrast enhancement using methods like CLAHE to improve the visibility of nodules (Gupta *et al*., 2023; Kamath & Singh, 2024). To address the everyday challenges of small dataset sizes and class imbalance, data augmentation is a near-universal

practice (Anand *et al*., 2022; Jassim *et al*., 2024; Tandon *et al*., 2022). Some studies have further utilized oversampling techniques like the Synthetic Minority Oversampling Technique (SMOTE) to balance class distributions (Kumaran *et al*., 2024).

Architectural Models Commonly Applied in Lung Cancer Diagnosis:

The choice of model architecture is central to performance. We can categorize the approaches in the literature into three main groups:

1. Transfer Learning with Standard Architectures: The most prevalent approach involves leveraging pre-trained CNNs. Models such as VGG16 (Anand *et al*., 2022; Benamara *et al.,* 2024; Ghosh *et al*., 2023; Klangbunrueang *et al.*, 2025), MobileNetV2 (Alheeti *et al*., 2024; Ghosh *et al*., 2023), ResNet (Al-Shouka & Alheeti, 2023; Sangeetha *et al*., 2023), and InceptionV3 (Anand *et al*., 2022) are commonly fine-tuned for lung cancer classification, often achieving high accuracy scores. These studies have indicated the power of transfer learning in medical imaging.
2. Complex and Ensemble Models: To further boost accuracy, some researchers have developed more complex systems. This includes creating novel hybrid architectures, such as VCNet, which combines VGG16 with a Capsule Network (Tandon *et al*., 2022), or IRRCNN, which integrates Inception and recurrent layers (Anusha & Reddy, 2023). Another popular strategy is ensembling, where predictions from multiple models (e.g., ResNet, EfficientNet) are combined to produce a more robust final decision (Jassim *et al*., 2024; Kumaran *et al*., 2024). These approaches often report state-of-the-art accuracies, frequently exceeding 99% on their respective test sets.
3. Handcrafted CNNs: A smaller subset of studies builds custom, handcrafted CNNs from scratch (Anand *et al*., 2022; Gupta *et al*., 2023). While sometimes effective, these methods are slightly outperformed by transfer learning and can be less scalable.

Validation Strategies and Generalization:

A model's true clinical value is determined by its ability to generalize to new, unseen data. However, the validation strategies reported in the literature are often limited. The vast majority of studies evaluate their models using a simple train-test split on a single, homogenous dataset, such as LIDC-IDRI or IQ-OTH/NCCD (Kumaran *et al*., 2024; Tandon *et al*., 2022; Ghosh *et al*., 2023; Gupta *et al*., 2023; Naseer *et al*., 2023).

Clinical Interpretability:

For a diagnostic AI tool to be trusted by clinicians, it must be interpretable. Explainable AI (XAI) techniques like Grad-CAM, which generate heatmaps to show where the model is focusing, are crucial for building this trust. However, XAI is still underutilized in the field, with only a few recent studies incorporating it (Kumaran *et al*., 2024; Klangbunrueang *et al*., 2025).

Limitations in Existing Literature and Our Proposed Solutions:

Although deep learning has advanced lung cancer classification, a review of the literature reveals several persistent limitations that hinder the development of clinically robust models. The present study directly addresses these challenges related to data preparation, model complexity, and validation.

First, many studies employ inadequate preprocessing methods without quantitatively validating their impact. For instance, some rely on basic filters alone (Ravindra *et al.*, 2024; Sangeetha *et al*., 2023), while others apply standard contrast enhancement techniques (Alheeti *et al*., 2024; Gugulothu & Balaji, 2023). Similarly, data augmentation is frequently simplistic (Anand *et al*., 2022; Tandon *et al*., 2022), and oversampling methods like SMOTE are often used in isolation, overlooking the need for greater data diversity (Kumaran *et al*., 2024). In contrast, our framework implements an advanced pipeline combining a hybrid filter with dynamic contrast enhancement to improve image quality. Furthermore, we pair advanced augmentation—designed to simulate real-world variations from different CT scanners and imaging protocols—with SMOTE to ensure the model trains on a diverse and balanced dataset.

Additionally, other research often relies on overly complex architectures to achieve high accuracy. These ensemble or deep models typically have high computational costs and memory requirements (Anusha & Reddy, 2023; Jassim *et al*., 2024; Tandon *et al*., 2022). Consequently, their use is impractical in many resource-constrained clinical settings.

To address this, our work streamlines the VGG16 architecture through structured pruning. This method reduces the model's parameters by approximately 80%, creating a lightweight yet powerful model suitable for real-world deployment.

Furthermore, a critical weakness in the literature is the reliance on limited validation strategies. Most studies evaluate their models on a single data source with a basic train-test split (Alheeti *et al*., 2024; Al-Shouka & Alheeti, 2023; Balaji & Gugulothu, 2023; Jassim *et al*., 2024; Kumaran *et al*., 2024; Klangbunrueang *et al.*, 2025; Naseer *et al*., 2023;); consequently, these models often fail to prove they can generalize to different clinical environments. Performance is also often measured with narrow metrics like accuracy (Tandon *et al*., 2022; Sangeetha *et al*., 2023), which can be misleading.

Our study directly confronts these issues with a multi-layered validation strategy. First, we apply K-Fold Cross-Validation on the single-source dataset. Second, and most importantly, we retrained and validated our entire framework on a large, multi-source dataset—a step notably absent in much of the literature—to confirm its real-world generalization. Both validation approaches are evaluated with comprehensive metrics, including F1-score, MCC, 95% CI, Cohen's Kappa, and error-based metrics.

Moreover, explainability is underutilized, with only rare studies implementing it (Kumaran *et al*., 2024; Klangbunrueang *et al*., 2025). We address this gap by integrating Grad-CAM, which provides transparent visualizations to help build clinical trust.

3: METHODS AND MATERIALS

This section outlines the experimental methodology developed for our automated lung cancer detection framework. Our end-to-end pipeline integrates a quantitatively validated preprocessing stage, a hybrid data balancing strategy, a pruned and fine-tuned VGG16 model, a multi-layered validation protocol, and Grad-CAM for interpretability. This approach was systematically designed to overcome common limitations in the literature, including poor data quality, model complexity, and

limited generalization. The overall workflow is presented in Figure 1, and each step is detailed in the sections below.

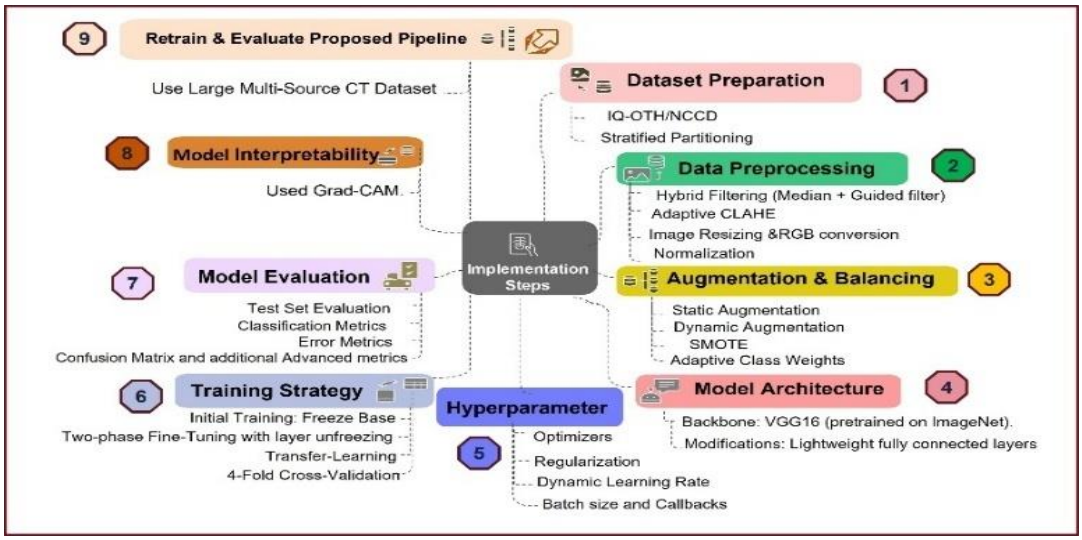


Figure 1: Implementation Steps of Proposed Methodology

Dataset Description:

The proposed model was developed and validated using the publicly available IQ-OTH/NCCD dataset (Al-Yasriy *et al*., 2020). This dataset, professionally annotated by expert radiologists and oncologists, comprises 1,097 grayscale CT images, each with a resolution of 512×512 pixels. The images are categorized into three clinically relevant classes: Benign, Malignant, and Normal. As stated in Table 1, the dataset exhibits two primary challenges: a modest overall sample size and a significant class imbalance, with malignant cases constituting over half of the data. These characteristics necessitate a rigorous data partitioning and validation strategy to ensure a robust and unbiased model evaluation.

Table 1: Class Distribution of the IQ-OTH/NCCD Dataset

Class Type	Description	Number of Samples	Percentage (%)
Benign	Non-cancerous lung nodules	120	10.94%
Malignant	Cancerous (lung cancer) nodules	561	51.14%
Normal	Healthy lung images with no nodules	416	37.91%
Total	—	**1,097**	**100%**

Preprocessing Pipeline and Comparative Evaluation:

Effective preprocessing is fundamental to enhancing the quality and diagnostic utility of lung CT images for deep learning models. This study, therefore, systematically investigates various preprocessing techniques to identify an optimal pipeline for lung nodule classification. The goal is to address key artifacts in medical images, specifically noise, contrast variability, and spatial inconsistencies, by evaluating the performance of multiple denoising and contrast enhancement methods. Quantitative evaluation was performed using two widely accepted metrics, the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM). PSNR measures pixel-level fidelity, while SSIM provides a more perceptually relevant assessment by comparing luminance, contrast, and structural information (Al Najjar, 2024; Rodrigues *et al*., 2024). Higher values for both metrics signify superior image quality and structural preservation. A detailed summary of the preprocessing configurations and their parameters is provided in Table 2.

Table 2: Summary of Preprocessing Steps and Parameters Setup

#	Step	Technique	Parameters	Purpose
1	Noise Reduction	Hybrid Filtering (Median + Guided)	- Median Filter kernel 3×3 - Guided Filter (radius=5, smoothness controller ε=0.1)	Removes salt-and-pepper noise while preserving critical edge and texture details, outperforming single-filter methods.
2	Contrast Enhancement	Dynamic CLAHE	- Tile Grid: 8×8 - Base Clip Limit: 1.5 - Scale Factor: 0.3 - Entropy Thresholds: E_{Low} =5, E_{High}=6	Enhances local contrast for better feature visibility while preventing noise over-amplification in high-contrast regions.
3	Standardization	Resizing & Normalization	- Method: Lanczos Interpolation - Target Size: 224×224, Range: [0, 1]	Ensures uniform input dimensions and pixel value range for stable training.

Hybrid Filtering for Denoising: Comparative Evaluation:

To suppress imaging noise while preserving anatomical structures, we evaluated several denoising algorithms to identify the most suitable technique. Table 3 summarizes the quantitative results across methods.

Table 3: Comparative Analysis of Denoising Techniques **(**Higher values indicate better quality)

Denoising Technique	PSNR (dB)	SSIM	Observations
Bilateral Filtering	30.75	0.7754	Over-smoothed; loss of texture
Gaussian Filtering	35.79	0.8729	Smooth output; edge blurring observed
Median Filtering (3×3)	39.82	0.8845	Strong denoising; moderate structure preservation
Median Filtering (5×5)	36.28	0.8191	Excessive smoothing; reduced fine detail retention
Guided Filtering	36.54	0.8992	Highest SSIM; preserves edges well, less effective on noise.
Non-Local Means (NL-Means)	36.86	0.8595	Good PSNR, slight blurring; time-consuming
NL-Means + Guided Filtering	35.73	0.8298	Redundant smoothing; degraded structure clarity
NL-Means + Bilateral Filtering	29.53	0.6934	Suboptimal; lowest PSNR and SSIM
Proposed: Median (3×3) + Guided	**39.88**	**0.8845**	**Best PSNR, high SSIM; balanced edge/detail preservation**

As shown in Table 3, the Hybrid Filtering method (combining Median 3×3 and Guided Filter) was selected as the optimal denoising strategy. It outperformed standalone techniques by achieving the highest PSNR (39.88 dB) and a strong SSIM (0.8845), effectively balancing noise reduction with structural preservation. This was achieved by effectively combining a Median filter's spatial smoothing with a Guided Filter's edge-aware properties. While Guided Filtering alone yielded a higher SSIM (0.8992), its lower PSNR suggested undesirable over-smoothing. Therefore, the hybrid configuration was implemented as the foundation for our preprocessing pipeline.

Following denoising, local contrast enhancement was performed using Contrast-Limited Adaptive Histogram Equalization (CLAHE) to improve the visibility of fine structures in lung CT images. Unlike standard CLAHE, which uses a fixed clip limit and may result in over-amplification of noise in homogeneous or high-contrast regions, we implemented a dynamic clip limit adjustment based on image entropy. This strategy is designed to prevent over-enhancement in high-contrast images while selectively boosting contrast where it is most needed. The adaptive clip limit is computed using the following Equation (1):

Dynamic CLAHE for Contrast Enhancement:

$$CL_{dynamic} = \begin{cases} CL_{base} + \left(SF \times (E_{Low} - Entropy)\right) & if\ E < E_{low} \\ CL_{base} - \left(SF \times \left(Entropy - E_{High}\right)\right) & if\ E_{High} > E \\ CL_{base} & otherwise \end{cases} \quad \ldots\ldots\ldots\ldots\ldots (1)$$

Where *Entropy* represents image histogram entropy. CL_{base} (1.5) is the default base clip limit; *SF* (default: 0.3) is the scaling factor, while E_{Low} (5) and E_{High} (6) are the entropy thresholds. The adjustment mechanism is controlled by these entropy thresholds. For images with entropy exceeding the upper threshold of 6, the default base clip limit is applied. This is a thoughtful design choice, as the entropy values in our dataset ranged from 4.98 to 7.12 (mean = 6.20). By setting the threshold slightly below the average, we ensure that the majority of images are treated cautiously, minimizing the risk of noise amplification and structural artifacts in these already detailed images. In contrast, for the rare images with entropy below the lower threshold of 5, the clip limit is increased using a scaling factor (SF) to significantly boost contrast. This approach of using entropy to guide the clip limit aligns with the recommendations of Chang *et al.* (2018), who demonstrated its benefits for medical image enhancement.

The effectiveness of our proposed dynamic CLAHE was evaluated against the standard approach using PSNR and SSIM metrics. As reported in Table 4, the adaptive method achieved higher PSNR and markedly superior SSIM than the fixed-limit baseline, confirming better structural preservation and perceptual quality.

Table 4: Performance Comparison of Contrast Enhancement Techniques

Contrast Enhancement Technique	Average PSNR (dB)	Average SSIM
Dynamic CLAHE (Adaptive Clip Limit, 8×8 Tile Grid)	31.08	0.8584
Standard CLAHE (Fixed Clip Limit: 1.5, 8×8 Tile Grid)	29.12	0.7288

The effect of the proposed preprocessing pipeline on CT lung images, along with corresponding histogram visualizations, is illustrated in Figure 2. As observed, the hybrid filtering stage effectively reduces background noise while preserving critical anatomical structures. Following this, the application of dynamic CLAHE significantly enhances local contrast, improving the visibility of subtle pathological features. The accompanying histograms reveal a notable redistribution and widening of pixel intensity values, confirming enhanced feature differentiation and improved perceptual quality.

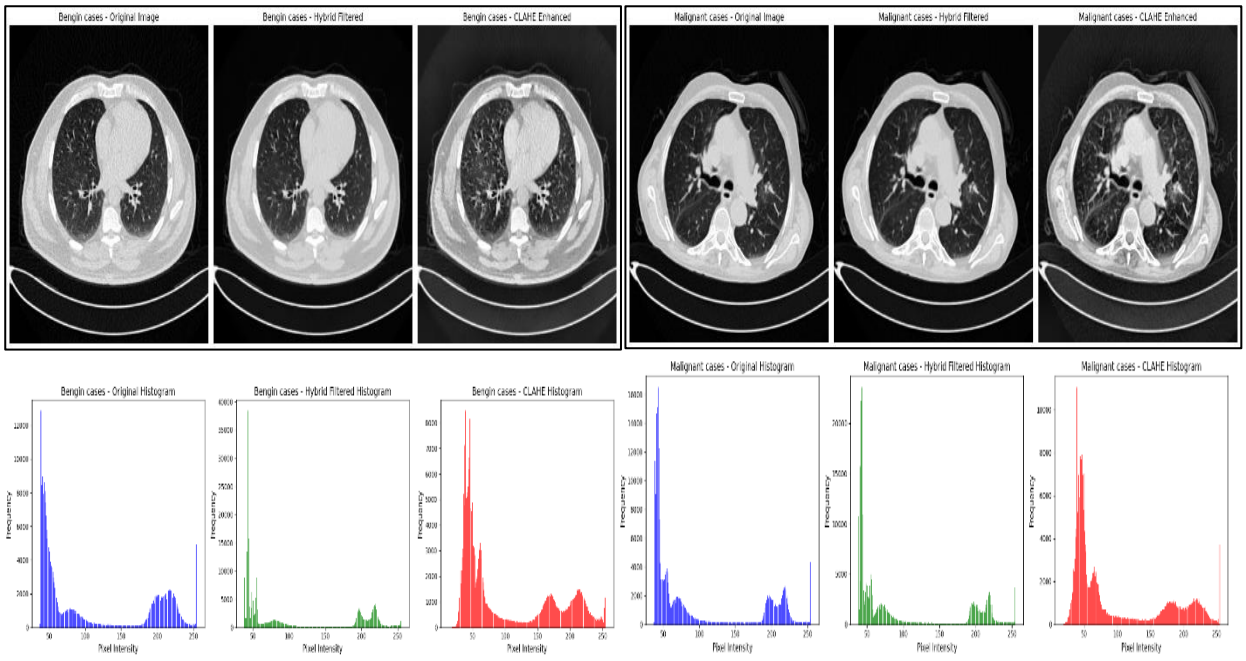


Figure 2: Effect of Preprocessing on CT Lung Images with Histogram Visualization
The figure illustrates representative CT scan samples at different stages of preprocessing. From left to right: (1) Original grayscale image, (2) Denoised output using Hybrid Filtering (Median + Guided), and (3) Contrast-enhanced image using dynamic CLAHE. Each image is accompanied by its corresponding histogram to highlight intensity distribution changes.

Hybrid Data Augmentation and Balancing Strategy:

To address the limited dataset size and pronounced class imbalance, a comprehensive, multi-stage data augmentation and balancing strategy was employed within each training fold of the cross-validation. This hybrid approach was designed to enhance data diversity and mitigate class underrepresentation, thereby improving model generalization, as detailed below.

Static (Offline) Augmentation:

Before training started, the images in the training set underwent offline augmentation. This involved applying a set of transformations, including affine transformations, elastic deformations, and cut-out augmentation, to the original images. The augmentation rate was class-specific to target underrepresented classes more aggressively. The number of augmented images for each class was determined by Equation 2:

$$N_{aug} = N_{original} \times \max\left(min\_aug, \frac{1000}{N_{original}}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Where N_{aug} is the total number of augmented images for the class, $N_{original}$ is the count of original images, and min_aug It is the minimum augmentation multiplier. The threshold of 1,000 images per class was established as a robust baseline for balancing to ensure that even the most underrepresented classes (Benign) had a sufficiently large and diverse set of samples for the model to learn meaningful features. This value was determined through preliminary experiments, which indicated that smaller thresholds led to underfitting on the minority classes, while significantly larger thresholds offered diminishing returns on performance at a higher computational cost. This approach provided a balanced trade-off between data diversity and training efficiency. Furthermore, multiprocessing was used to parallelize transformations and improve efficiency.

Dynamic (Real-time) Augmentation:

During the training process, on-the-fly transformations were applied to each batch of data using the Keras Image Data Generator. These included random rotations (±10°), brightness variations (scale 0.7–1.3), zooming, width/height shifts (10%), and shear distortions. This dynamic approach ensures the model encounters slightly different versions of the images in each epoch, which is highly effective at reducing overfitting and improving model robustness without requiring additional storage.

Feature-Space Synthetic Oversampling:

To further address class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied in the feature space rather than on raw pixel data. During training, CT images were passed through the convolutional base of the VGG16 model to extract high-level embeddings, and SMOTE was used to generate synthetic feature vectors for minority classes by interpolating within this learned feature space. This approach preserves semantic consistency and avoids the visual artifacts often introduced by pixel-level oversampling, thereby enhancing minority class representation while maintaining clinical plausibility. Figure 3 shows the effect of the hybrid augmentation and balancing pipeline on the training set's class distribution.

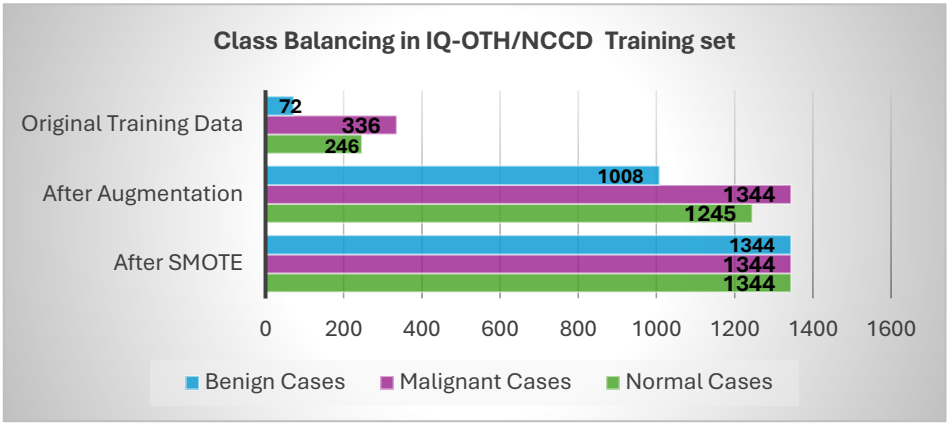


Figure 3: Class Distribution in the Training Set After Multi-Stage Balancing
The original imbalance (72 benign, 246 normal, 336 malignant) was mitigated through class-specific augmentation, followed by SMOTE, resulting in 1,344 samples per class. This improved diversity and generalization, particularly for the minority class.

Dynamic Class Weighting:

To mitigate residual class imbalance during training, dynamic class weighting was employed. This approach was especially important within the K-fold cross-validation framework, where stratified sampling maintains global class proportions but may still produce imbalances within individual batches. By recalculating class weights dynamically for each batch based on its class distribution, the model gives more importance to mistakes made on minority class samples. This real-time adjustment ensures consistent focus on underrepresented classes, enhancing feature learning and improving overall model fairness. Combined with augmentation and feature-space oversampling, this strategy forms a robust and generalizable learning pipeline.

Validation Strategy:

To ensure a comprehensive and robust evaluation, we employed a multi-layered validation strategy. This approach includes both a rigorous internal validation on our primary dataset and a crucial external validation on a large, multi-source dataset to confirm real-world generalization.

Data Partitioning and Internal Validation Strategy on the Primary Dataset:

Our internal validation strategy was designed to ensure robust model training, stable hyperparameter tuning, and an unbiased final evaluation. This was achieved through a multi-stage process involving an initial data split, followed by K-fold cross-validation on the training portion.

1. Initial Data Partitioning

First, the entire IQ-OTH/NCCD dataset was partitioned using stratified sampling into three distinct, non-overlapping subsets: Training Set (60%), reserved exclusively for training the model using a cross-validation protocol. Fixed Validation Set (20%), a hold-out set used as a consistent benchmark across all training iterations for early stopping and model checkpointing. Hold-Out Test Set (20%), a completely separate set used only once for the final, unbiased performance evaluation of the best model.

The detailed distribution of samples across these splits is presented in Table 5.

Table 5: Stratified Train-Validation-Test Split of the IQ-OTH/NCCD Dataset

Set	Benign	Malignant	Normal	Total
Training Set (for CV)	72	336	249	657
Fixed Validation Set	24	112	83	219
Hold-Out Test Set	24	113	84	221

K-Fold Cross-Validation Protocol:

To train the model robustly, we applied a 4-fold cross-validation (CV) protocol exclusively to the Training Set (the 60% portion). In each of the four training iterations, three folds were used for model training, while the remaining fold served as an internal validation set to provide immediate feedback on generalization within the training data.

Model Selection and Final Evaluation:

Crucially, during each training iteration, model performance was monitored on the Fixed Validation Set (the 20% hold-out). This set provided a stable, consistent benchmark across all four CV runs. Decisions for early stopping and saving the best model checkpoint were based exclusively on the performance on this Fixed Validation Set, ensuring that the model selection process was stable and not influenced by the variability of the individual validation folds.

Finally, the single best-performing model identified through this entire process was evaluated once on the Hold-Out Test Set to report its final, unbiased performance.

Generalization Assessment on Multi-Source Dataset:

To assess the model's real-world generalization and robustness beyond a single data source, a final external validation was performed. This validation utilized a large, composite multi-source dataset comprising over 29,000 images compiled from

five public sources, including our primary IQ-OTH/NCCD dataset.

For this assessment, the entire proposed framework—from the preprocessing pipeline to the model training and fine-tuning strategy—was reapplied to this new, diverse dataset. The model, after training, was then evaluated using a standard train-test split to simulate a realistic deployment scenario. This critical step confirms that the model's high performance is not limited to the characteristics of a single dataset and that it can generalize effectively to unseen data from different clinical environments. The specific composition of this dataset and the detailed outcomes of this evaluation are presented in Section 4.4.

Proposed Model Architecture and Efficiency Improvement:

The core of our framework is the pretrained VGG16 model, which consists of 13 convolutional layers grouped into 5 blocks (indexed from block1_conv1 through block5_conv3), followed by three large fully connected (FC) layers. While this structure provides excellent feature extraction, its original design is too large and computationally intensive for practical clinical use. Therefore, we implemented a structured, layer-level pruning strategy. Specifically, we removed the original, oversized FC layers in their entirety, including all associated weights and neurons. This was necessary because these layers are a primary source of computational cost and are highly prone to overfitting, a problem that is worsened when working with modest-sized medical datasets. Figure 4 illustrates architectural differences between the original and modified model.

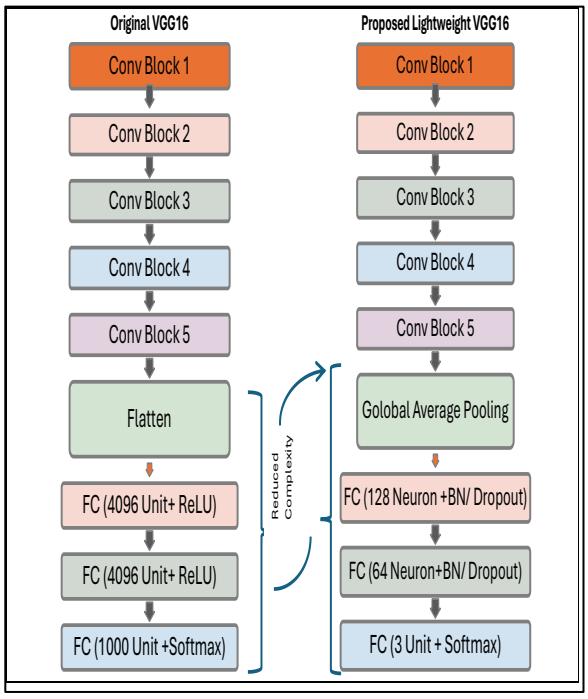


Figure 4: Structural Comparison Between the Original VGG16 and the Proposed Lightweight VGG16 Model

Therefore, the resulting lightweight architecture significantly reduces the model's size by approximately 80%, making it over five times smaller than the original VGG16. This promotes efficient deployment of our model while simultaneously improving predictive accuracy by enhancing generalization. Figure 5 presents a comparison of parameters and model size between the original and modified versions.

Figure 5: Comparison of Parameter Count and Model Size between the Original VGG16 and the Modified Version.

Training Strategy:

Transfer Learning and Fine-Tuning:

To effectively utilize transfer learning, we adopted a multi-stage training strategy based on a modified pretrained VGG16 architecture. The model was designed for multi-class classification using the Categorical Cross-Entropy loss function. The convolutional layers served as feature extractors, capturing essential patterns from lung CT images.

In the initial training phase, the model was trained for 30 epochs using the Adam optimizer, while the base VGG16 layers were frozen and only the newly added custom classification head was updated. This step preserved the general-purpose features learned from ImageNet and enabled the model to start learning lung-specific patterns without modifying the core feature extractors. Following this, we implemented a two-phase fine-tuning strategy to gradually adapt the model to the unique characteristics of lung CT images while retaining the general representations learned during pre-training:

- First Fine-Tuning Phase:

The model was trained for 20 additional epochs using the Stochastic Gradient Descent (SGD) optimizer with the last 10 convolutional layers unfrozen. These included layers from block3_conv1 to block5_conv3, which are responsible for extracting mid- and high-level semantic features. Unfreezing these layers allowed the model to fine-tune the deeper representations related to lung tissue and lesion structures.

- Second Fine-Tuning Phase:

To further refine feature learning, we unfroze five additional earlier layers—including block2_conv1, block2_conv2, and earlier layers in block3. The model was retrained for another 10 epochs using the same optimizer with a lower initial learning rate. This allowed more accurate weight updates in both mid- and high-level layers, improving convergence and reducing overfitting.

Eventually, this selective fine-tuning strategy, where layers were gradually unfrozen from deep to shallow, enabled effective adaptation of pretrained features to the medical imaging domain. It also ensured stable training dynamics and strong generalization performance on the multi-source lung CT dataset.

Adaptive Learning Rate Scheduling:

To ensure stable and efficient model convergence, we implemented a custom adaptive learning rate scheduling strategy. This approach is designed to mitigate the risks of divergence

during early training and to refine the model's parameters as it approaches an optimal solution. The schedule consists of two distinct phases: an initial warm-up phase followed by a stepwise decay phase. This two-phase process is detailed in Algorithm 1 and formulated in Equation 14.

Algorithm 1: Adaptive Learning Rate Schedule

Input: Current epoch E, initial learning rate lr_initial, minimum learning rate lr_min, number of warm-up epochs E_warmup, decay rate R_decay, epoch drop period E_drop.
Output: Calculated learning rate lr for the current epoch.

```
if E < E_warmup then
// Warm-up Phase: Linearly increase LR from a low value to lr_initial.
lr ← lr_initial × (E + 1) / E_warmup
else
// Decay Phase: Apply stepwise exponential decay.
epochs_since_warmup ← E - E_warmup
decay_steps ← floor(epochs_since_warmup / E_drop)
lr ← lr_initial × (R_decay ^ decay_steps)
end if
// Clamping: Ensure the learning rate does not fall below the minimum threshold.
lr ← max(lr, lr_min)
return lr
```

The mathematical formulation for the learning rate, lr(E), at a given epoch E is defined as:

$$l_r(E) = \begin{cases} lr_{initial} \times \frac{E+1}{E_{warmup}} & if\ E < E_{warmup} \\ \max\left(lr_{initial} \times R_{decay}^{\left\lfloor \frac{E - E_{warmup}}{E_{drop}} \right\rfloor}, lr_{min} \right) & if\ E \geq E_{warmup} \end{cases} \quad (14)$$

The description and values of parameters used in this schedule are defined in Table 6.

Table 6: Learning Rate Scheduler Parameters

Parameter	Symbol	Value	Description
Initial Learning Rate	$lr_{initial}$	0.001	The target learning rate is after the warm-up phase.
Minimum Learning Rate	lr_{min}	1e-6	The lower value for the learning rate is to prevent training from stalling.
Warm-up Epochs	E_{warmup}	5	The number of epochs for the linear warm-up phase.
Decay Rate	R_{decay}	0.75	The multiplicative factor for each decay step.
Epoch Drop	E_{drop}	5	The number of epochs between each learning rate decay.

This adaptive During the warm-up phase (the first 5 epochs), the learning rate increases linearly. This gradual ramp-up allows the model to stabilize by taking smaller, more cautious steps when the model's weights are still randomly initialized and gradients can be large and erratic. Following the warm-up, the schedule transitions to the stepwise decay phase. The learning rate is reduced by a factor of 0.75 every 5 epochs. This allows the model to make larger updates early in the decay phase and progressively smaller, more refined updates as it converges, helping to prevent overshooting the minima in the loss landscape. Finally, the learning rate is clamped at a minimum value (lr_{min}) to ensure that training does not halt prematurely. Figure 6 provides a visualization of this adaptive schedule, illustrating the linear increase during the warm-up period followed by the discrete, stepwise decay throughout the remainder of the training process.

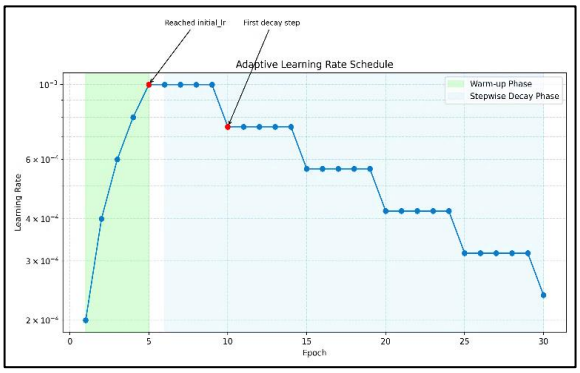


Figure 6: Learning Rate Path During Training
Visualization of the adaptive LR schedule. The process begins with a 5-epoch linear warm-up to a rate of 0.001, followed by a stepwise exponential decay. The discrete drops in the learning rate help refine the model's convergence as training progresses.

Early Stopping and Model Checkpointing:

Early stopping was employed by monitoring the loss on a fixed clean validation set to prevent overfitting. The model automatically reverted to the weights corresponding to the epoch with the lowest validation loss. During training, checkpoints were saved at the end of each epoch, but only the weights that achieved the minimum loss on the validation set were retained for final evaluation.

Performance Evaluation:

Evaluation Metrics:

To assess model performance robustly, a diverse set of metrics was used. Precision minimized false positives, while

recall ensured detection of actual cancer cases. The F1-score balanced these metrics; it is useful in the presence of class imbalance. AUC-ROC evaluated all class separability across thresholds. Cohen's Kappa and Quadratic Weighted Kappa (QWK) measured agreement beyond chance, and they are more reliable than accuracy alone, accounting for the clinical severity of misclassifications. MCC offered a balanced summary of all confusion matrix elements, and balanced accuracy ensured fair class-wise performance. While Error prediction metrics (MSE, RMSE, and MAE) are employed to provide a quantified prediction of deviations, the 3×3 confusion matrix visualizes per-class accuracy and guided refinement. ALL these metrics together supported a comprehensive, reliable, and interpretable evaluation of our proposed model. A full summary of the mathematical definitions of these metrics is presented in Table 7.

Table 7: Comprehensive Evaluation Metrics for Classification Performance

Metric	Mathematical Formulation of Metrics	Equation No.
Accuracy	$Acc. = \frac{TP + TN}{TP + TN + FP + FN}$	(3)
Precision	$Prec. = \frac{TP}{TP + FP}$	(4)
Recall	$Rec. = \frac{TP}{TP + FN}$	(5)
F1-score	$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$	(6)
Cohen's Kappa (κ)	$\kappa = \frac{p_o - p_e}{1 - p_e}$	(7)
Quadratic Weighted Kappa (QWK)	$QWK = 1 - \frac{\sum w_{ij} O_{ij}}{\sum w_{ij}\, E_{ij}}$	(8)
Matthews Correlation Coefficient (MCC)	$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$	(9)
Balanced Accuracy	$Balanced\ Acc. = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)$	(10)
Mean Squared Error (MSE)	$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	(11)
Root Mean Squared Error (RMSE)	$RMSE = \sqrt{MSE}$	(12)
Mean Absolute Error (MAE)	$MAE = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i \rvert)$	(13)

Abbreviations:

TP: True Positive, *TN*: True Negative, *FP*: False Positive, *FN*: False Negative, y_i: Actual value, $\hat{y}_i$: Predicted value, *n*: Number of samples, w_{ij}: Weight matrix, O_{ij}: Observed agreement matrix, E_{ij}: Expected agreement matrix, p_o: Observed agreement, p_e: Expected agreement.

Statistical Analysis:

The model's performance was assessed using a multi-layered evaluation strategy to ensure robustness and clinical relevance. The initial phase involved a rigorous internal validation on the primary single-source dataset, conducted via 4-fold cross-validation. Subsequently, a final generalization assessment was performed by retraining and validating the entire pipeline on a large, independent multi-source dataset. Throughout both stages, performance was quantified using a comprehensive suite of metrics, including F1-score, Matthews Correlation Coefficient (MCC), AUC, and 95% Confidence Intervals (CI).

Model Interpretability with Grad-CAM:

This study applied a Grad-CAM technique in order to improve the model's interpretability and ensure the model focuses on clinically relevant regions within the lungs. Grad-CAM relies on guided backpropagation to make the most important parts of the image visible and highlight them with heat maps (Chattopadhay *et al*., 2018). This visualization technique shows which parts of the image play a key role in the results, with red/yellow areas usually containing nodules or abnormal tissue in both malignant and benign cases, and blue regions often being background or normal lung tissue. It provides intuitive insights into the model's decision-making process and could potentially enhance its clinical reliability.

EXPERIMENTAL RESULTS AND EVALUATION

Experimental Setup:

To ensure efficient training and evaluation, the proposed framework was implemented in a high-performance computing environment. The experiments were conducted on Kaggle's cloud-based platform using dual NVIDIA T4 GPUs. The software environment was configured using Python, TensorFlow, and Keras, in addition to different essential libraries for model implementation.

Hyperparameter Setup:

A carefully selected set of hyperparameters was configured to ensure optimal performance and stability of the proposed model. These settings were chosen based on experimental validation and deep learning best practices to enhance

convergence efficiency, generalizability, and robustness. The hyperparameter configuration is stated in Table 8.

Table 8: VGG16 Hyperparameters Setup

Hyperparameter	**Value/Type**
Added Layers	2 Dense Layers
Neurons per Layer Activation	128, 64
Function Regularization	ReLU (both layers)
Batch Normalization	$\lambda = 0.001$ for L1, $\lambda = 0.01$ for L2
Dropout Rate	Applied after each dense layer
Output Layer	0.5 (after each dense layer)
Optimizer (Phase 1)	Softmax with 3 neurons (for 3-class classification)
Optimizer (Fine-Tuning)	Adam
Initial Learning Rate Learning	SGD (with momentum 0.9)
Rate Schedule	0.001
LR Callback	Linear warmup (5 epochs) + piecewise decay (×0.75 every 5 epochs)
Batch Size (Phase 1)	ReduceLROnPlateau (min LR = 1e-6)
Batch Size (Fine-Tuning)	32
Early Stopping	16
Loss Function	Patience = 5 (based on validation loss)
	Categorical cross-entropy

Results on Primary Single-Source Dataset:

Training Performance:

First, we trained the proposed model on a single-source dataset (IQ-OTH/NCCD) for a 60-epoch training cycle, divided into a 30-epoch initial phase followed by two fine-tuning phases of 20 and 10 epochs, respectively. To visualize the learning process and verify model stability, the training and validation dynamics were monitored throughout the multi-phase training scheme for each fold. Figure 7 illustrates the accuracy and loss paths, providing a clear view of the model's convergence and generalization behavior by tracking performance on the training set, an internal validation set, and a clean fixed validation set.

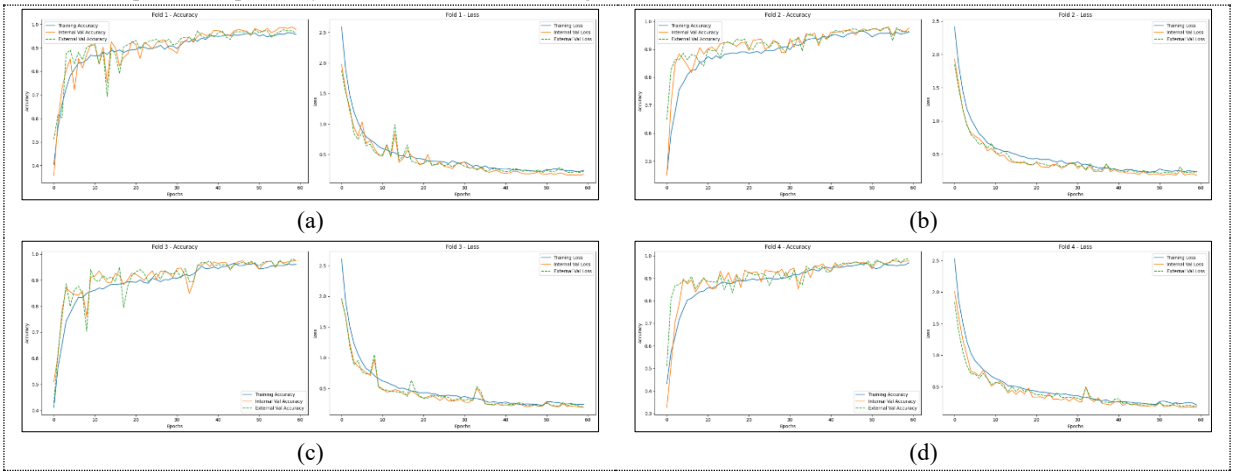


Figure 7: The Training and Validation Performance Across 4 Folds.
(a)–(d) show accuracy (top) and loss (bottom) for Folds 1–4. Curves represent the training set (augmented + SMOTE), internal validation (from training), and validation (clean, fixed subset) across three training phases.

The learning curves presented in Figure 7 provide strong evidence of a stable and effective training process. Across all four folds, there is a consistent and smooth increase in accuracy and a corresponding decrease in loss for all three data splits, indicating successful model convergence.

Validation and Test Performance:

To quantitatively assess the best performance model checkpoints, the maximum validation accuracies were recorded for both the internal (augmented and SMOTE-balanced) and fixed (clean) validation sets within each fold, as illustrated in Figure 8.

The performance on the validation set (fixed and clean) closely tracks the performance on the training and internal validation sets, with only a minimal gap between peak accuracies. This demonstrates that the model generalizes exceptionally well to unseen, clean data and is not overfitting to the augmented training distribution.

The progressive improvement across the initial training and the two-phase fine-tuning stages further validates the efficacy of the gradual unfreezing strategy, allowing the model to effectively adapt its learned features without destabilizing the training process.

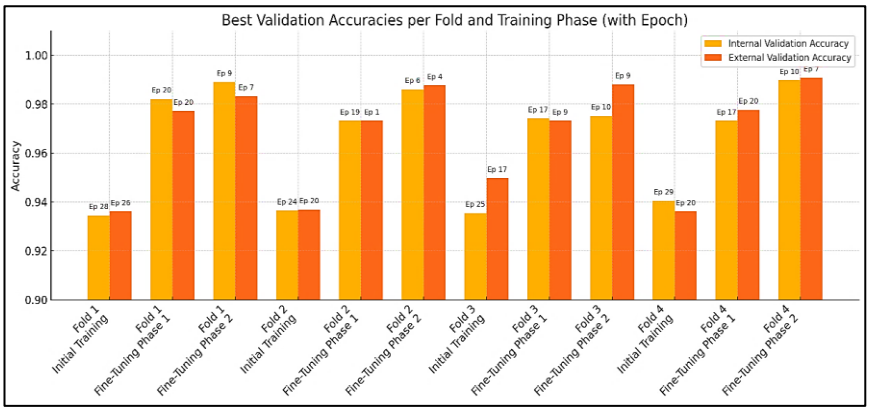


Figure 8: Peak Validation Accuracies Across Folds and Training Phases.
This figure compares the highest validation accuracies achieved on the internal (augmented and balanced) and fixed (clean) validation sets for each of the four folds. Each bar represents the peak accuracy reached during the 60-epoch training, and the number annotated on the bar indicates the specific epoch in that training phase where the maximum accuracy was recorded.

To comprehensively evaluate our proposed framework, the best-performing model from each fold was saved in (keras) format along with its corresponding weights. Upon completing training across all folds, each model was independently evaluated on the held-out test set using a comprehensive set of performance metrics. The results, including mean and standard deviation across all folds, are summarized in Table 9, providing insights into the model's effectiveness and stability.

Table 9: Detailed Performance Metrics of the Proposed VGG16 Framework Across 4-Fold Cross-Validation

Fold	Accuracy	F1 Score	Loss	MCC	Balanced Accuracy	Cohen's Kappa	QWK	AUC	MSE	RMSE	MAE
Fold 1	0.9910	0.9911	0.1605	0.9769	0.9583	0.9765	0.9765	0.9998	0.0543	0.2330	0.0271
Fold 2	0.9819	0.9821	0.1786	0.9465	0.9137	0.9450	0.9450	0.9909	0.1131	0.3363	0.0588
Fold 3	0.9910	0.9911	0.1705	0.9769	0.9693	0.9766	0.9766	0.9900	0.0407	0.2018	0.0226
Fold 4	0.9910	0.9911	0.1582	0.9845	0.9722	0.9844	0.9844	0.9890	0.0362	0.1903	0.0181
Mean ± SD	**0.9887 ± 0.0045**	**0.988± 0.0045**	**0.1670± 0.0082**	**0.971± 0.0159**	**0.9534 ± 0.0249**	**0.9706± 0.0166**	**0.970 ±0.016**	**0.9927± 0.0046**	**0.0611± 0.0334**	**0.2404± 0.0619**	**0.0317± 0.0177**

The results presented in Table 9 demonstrate the exceptional performance and robustness of the proposed framework. The model has achieved a mean accuracy of 0.9887 and a mean F1-score of 0.9888, indicating superior classification capability. Crucially, the low standard deviation across all metrics further confirms the model's high stability and consistent performance across different data partitions. However, while all folds performed exceptionally well, Fold 4 emerged as the optimal model, achieving the highest Matthews Correlation Coefficient (MCC) (0.9845) and the lowest error rates (MSE of 0.0362) and lowest loss (0.1582). This combination of high predictive accuracy and minimal error makes it the most reliable candidate for deployment, validating the model's potential for clinical application. Furthermore, to provide a more rigorous assessment of the model's stability and the certainty of its performance estimates, the 95% confidence intervals (CIs) were calculated for all key metrics across the four folds. These results are indicated in Table 10.

Table 10: 95% Confidence Intervals for Key Performance Metrics Across Folds
The intervals [Lower Bound – Upper Bound] represent the reasonable range for the true performance metric, providing insight into the model's statistical stability on each data partition (each fold).

Metric	Fold 1 [95% CI]	Fold 2 [95% CI]	Fold 3 [95% CI]	Fold 4 [95% CI]
Accuracy	[0.9683 – 1.0]	[0.9412 – 0.9910]	[0.9683 – 1.0]	[0.9683 – 1.0]
F1 Score (Macro)	[0.9336 – 1.0]	[0.8798 – 0.9774]	[0.9472 – 1.0]	[0.9336 – 1.0]
Balanced Accuracy	[0.9067 – 1.0]	[0.8472 – 0.9667]	[0.9270 – 1.0]	[0.9067 – 1.0]
AUC (OVR)	[0.9976 – 1.0]	[0.9964 – 1.0]	[0.9995 – 1.0]	[0.9978 – 1.0]
MCC	[0.9482 – 1.0]	[0.9041 – 0.9836]	[0.9470 – 1.0]	[0.9482 – 1.0]
Cohen's Kappa	[0.9465 – 1.0]	[0.8991 – 0.9835]	[0.9457 – 1.0]	[0.9465 – 1.0]
Log Loss	[0.0251 – 0.0842]	[0.0334 – 0.1456]	[0.0203 – 0.0632]	[0.0168 – 0.0706]
MSE	[0.0 – 0.1267]	[0.0362 – 0.2128]	[0.0 – 0.0995]	[0.0 – 0.1267]

The confidence intervals in Table 10 provide a precise view of the model's statistical reliability. For most validation folds, the intervals are consistently narrow and have high lower bounds. This suggests the model performs with high reliability. Furthermore, the upper bound often reached 1.0, indicating that the model is capable of near-perfect classification. In contrast, Fold 2 showed wider confidence intervals and lower performance bounds. This highlights that the data partition contained more challenging instances. However, the model's performance remained excellent even on this difficult fold, reinforcing its overall robustness. Additionally, the confidence intervals for error metrics were tightly clustered near zero, confirming low prediction error with high statistical confidence. In summary, this analysis demonstrates that the model is not only high-performing but also statistically stable, maintaining an excellent performance profile even under challenging conditions.

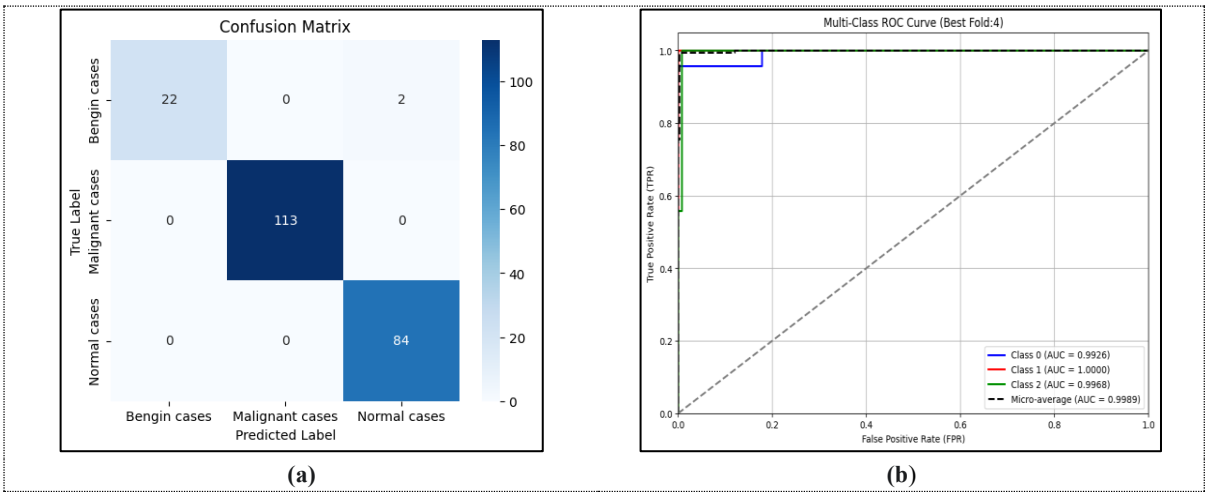
Confusion Matrix, Classification Report, and AUC-ROC:

The classification of the best model on the test set reported in Table 11 demonstrates the model's strong performance. Notably, the model achieved perfect precision, recall, and F1 scores for malignant cases. Performance on normal tissues was also excellent. However, the recall for benign cases was slightly lower, while precision remained perfect.

Furthermore, the confusion matrix in Figure 9 visually confirms these results. It shows that the only misclassifications were two benign cases incorrectly identified as normal. Finally, the Receiver Operating Characteristic (ROC) curve in Figure 9.b indicates outstanding diagnostic ability. Its near-perfect Area Under the Curve (AUC) score signifies high sensitivity and an extremely low false-positive rate. In summary, these metrics collectively validate the model's high accuracy and reliability, especially for identifying critical malignant cases.

Table 11: Classification Metrics of Best Fold Model (Fold 4)

Class	Precision	Recall	F1-Score	Support
Benign (0)	1.00	0.92	0.96	24
Malignant (1)	1.00	1.00	1.00	113
Normal (3)	0.98	1.00	0.99	84
Weighted Accuracy	**0.99**	**0.99**	**0.99**	**221**



(a) **(b)**

Figure 9:(a) Confusion matrix showing the proposed model classification performance on the test set (Fold 4). (b) ROC curves for Fold 4 showing AUC values of three classes reflecting superior discrimination capability across classes.

Furthermore, Figure 10 illustrates the model's predicted probability distribution for each class, visually representing its confidence in the classification decisions. This probability-based visualization enhances interpretability and reinforces the model's clinical reliability in distinguishing between different lung tissue types.

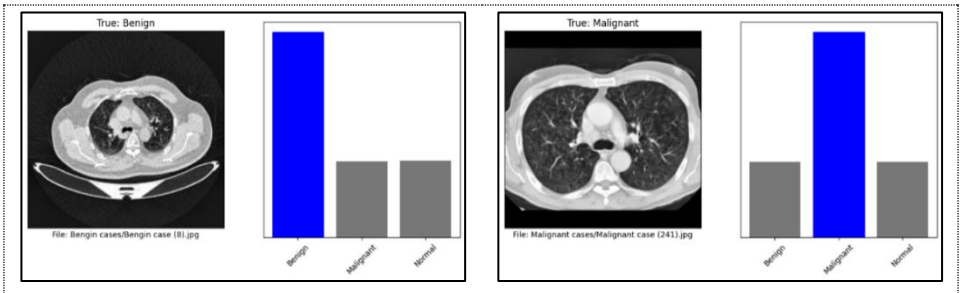


Figure 10: Samples of Probability Distribution Plot Among the Three Classes

Grad-CAM Visualizations Analysis:

The proposed VGG16 model effectively localized the relevant regions associated with abnormalities, focusing on areas consistent with potential tumor presence, as shown in Figure 11. Although not always perfectly aligned with the ground truth annotations, these head maps demonstrate that the VGG16 model captures significant discriminative features linked to malignant characteristics offers potential guidance for clinical assessment.

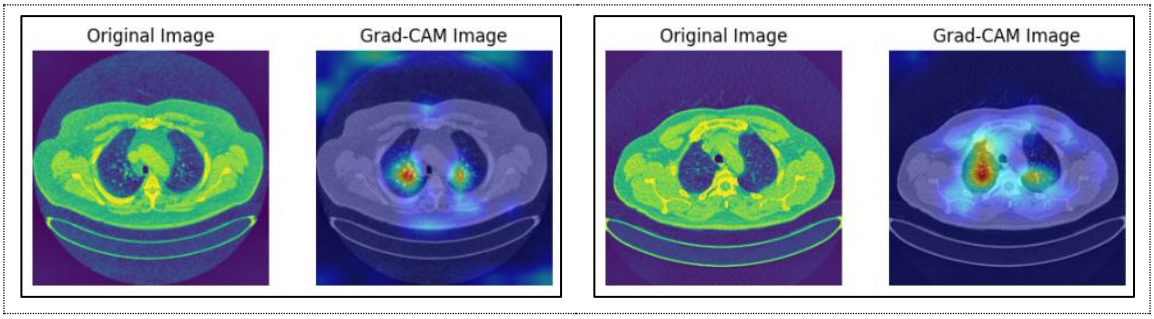


Figure 11: Grad CAM Visualization

Generalization to Multi-Source Datasets:

While the proposed model performed exceptionally well during 4-fold cross-validation on the primary IQ-OTH/NCCD dataset, relying on a single data source can limit a model's proven generalizability. To address the critical issue of domain shift (where models often fail on data from new clinical environments), we conducted a rigorous generalization assessment. Therefore, this evaluation aims to verify the model's capacity to generalize across varied imaging protocols, scanner types, and patient populations, which is a critical requirement for real-world deployment.

Dataset Composition and Label Harmonization:

To assess the model's generalization capability, we constructed a multi-source dataset. As detailed in Table 12, this composite dataset was formed by integrating five public lung CT datasets, resulting in a total of 29,546 images.

Table 12: Summary of Characteristics for the Unified Multi-source Lung CT Dataset

Source Dataset	Original Labels	Mapped Labels	Format	Image Count	Labeling Notes
IQ-OTH/NCCD	Normal, Benign, Malignant	Normal, Benign, Malignant	JPEG	1,097	Labels already follow the unified three-class scheme
Chest CT-Scan (Kaggle)	Normal; Adenocarcinoma, Large cell carcinoma, Squamous cell carcinoma	Normal, Malignant	PNG / JPEG	1,000	Malignant histological subtypes were grouped under "Malignant"
SPIE-AAPM-NCI L	Nodules (Benign), Nodules (Malignant)	Benign, Malignant	DICOM	15,931	Nodules manually labeled by radiologists and pathologically validated
Lung-RADS Dataset (Mendeley)	LR2, LR3 (Benign Appearance); LR4A (Suspicious), LR4B (Very Suspicious)	Benign, Malignant	PICKLE	972	Labels interpreted per Lung-RADS categorization scheme
LIDC-IDRI Subset	No nodules; Nodules with low rating malignancy; Nodules with high rating or confirmed malignancy	Normal, Benign, Malignant	DICOM	10,546	Malignancy labels derived from radiologist consensus and pathology reports
Final Unified Dataset	**--**	**Normal, Benign, Malignant**	**JPEG**	**29,546**	**All original labels harmonized into a unified 3-class format**

Before beginning any experiments on this dataset, a critical initial step involved a comprehensive data harmonization process. This began with mapping all original labels into a unified three-class scheme: Normal, Benign, and Malignant. Next, all images were standardized to 8-bit grayscale format using a fixed lung window (Level: 600 HU, Width: 1500 HU) to normalize intensity values across varying acquisition protocols. This harmonization process led to a clinically realistic class imbalance, with the final distribution as follows: Malignant with 13,656 images (46.2%), Benign with 11,490 images (38.9%), and Normal with 4,400 images (14.9%).

Re-training and Evaluation Results:

Following unification, our complete proposed methodology was applied. The dataset was first split into training (70%), validation (20%), and test (10%) subsets using stratified sampling. Our established pipeline, including hybrid preprocessing, class-specific data augmentation, and SMOTE-based balancing, was then applied exclusively to the training set. Figure 12 shows class distribution (Benign, Malignant, Normal) before and after augmentation and SMOTE in the training set.

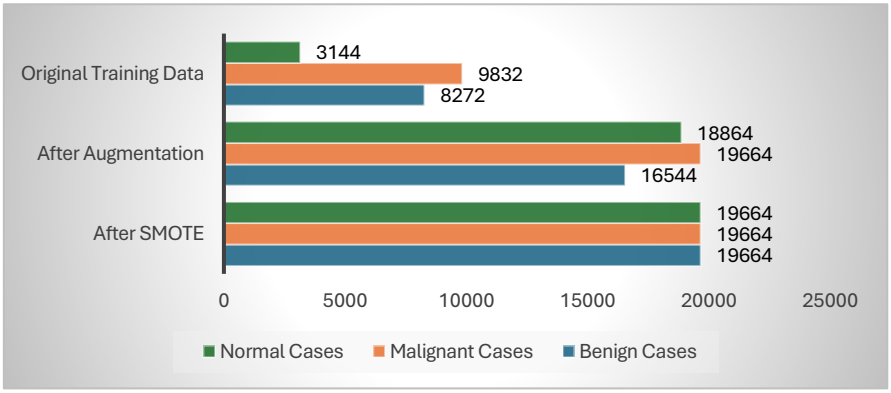


Figure 12: Class Distribution Before and After Balancing in The Multi-Source Training Set.

Following that, the proposed model was retrained and evaluated on the held-out test set. A single train-test evaluation was selected instead of cross-validation, given the dataset's large size and heterogeneity, offering a robust measure of the model's real-world generalizability. The training dynamics plots are illustrated in Figure 13.

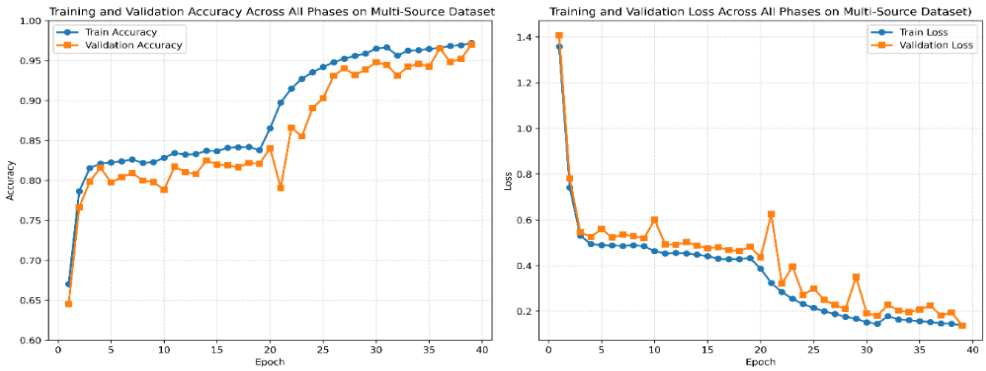


Figure 13: Training and Validation Accuracy (Left) and Loss (Right) Curves Across All Training Phases on The Multi-Source Dataset

Performance Metrics and Comparison:

When we evaluated on the held-out multi-source test set, the model demonstrated excellent generalization capabilities. The framework maintained high performance, confirming its ability to adapt to data from unseen sources. Key evaluation metrics, including agreement metrics and error-based metrics, are summarized in Table 13.

Table 13: Performance of Proposed Model on the Multi-Source Test Set

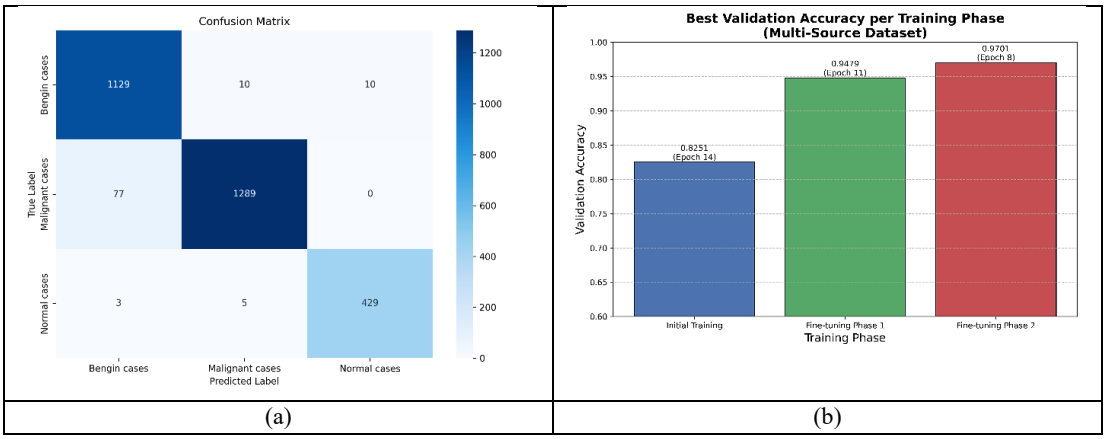
Metric	Performance Score
Balanced Accuracy	0.9693
Loss	0.1503
AUC	0.9980
MCC	0.9427
Cohen's Kappa	0.9421
MSE	0.0193
RMSE	0.1390
MAE	0.0440
QWK	0.9497

The results in Table 13 show high predictive accuracy, strong agreement scores, and low error values, indicating robust generalization. Additional classification metrics are presented in Table 14. The consistently strong performance of our proposed model, even on a challenging and heterogeneous dataset, suggests that it did not overfit to a single data source. Instead, it effectively learned generalizable radiological features indicative of lung pathology. These findings strongly support the model's robustness and highlight its potential for reliable deployment in real-world clinical environments with diverse scanners and patient populations.

Table 14: Classification Metrics on Multi-Source Test Set

Class	Precision	Recall	F1-Score	Support
Benign cases	0.93	0.98	0.96	1,149
Malignant cases	0.99	0.94	0.97	1,366
Normal cases	0.98	0.98	0.98	437
Accuracy			0.96	2,952
Macro avg	0.97	0.97	0.97	2,952
Weighted avg	0.97	0.96	0.96	2,952

Furthermore, Figure 14.b demonstrates a consistent increase in validation accuracy across training phases (initial training, two-phase fine-tuning), confirming the efficacy of selective unfreezing. This aligns with the strong classification performance in Figure 13. a: 1129 benign and 1289 malignant cases were correctly identified, while normal cases achieved 429 accurate predictions with minimal misclassifications (3 benign, 5 malignant). Critically, zero malignant cases were mislabeled as normal, underscoring the model's reliability. Together, these results highlight robust generalization on diverse data.



(a) (b)

Figure 14: (a) Final Confusion Matrix on Multi-Source Test Set. (b) Phase-Wise Best Validation Accuracy

4. DISCUSSION

This section interprets the experimental results presented in Section 4, connecting them to the methodologies detailed in Section 3. We discuss the significance of our findings and compare our framework's performance to the existing methods, highlighting its key advantages.

Summary and Interpretation of Findings:

The present study has successfully developed a lightweight, accurate, and robust framework for lung cancer classification. The outstanding performance reported in Section 4—including a test accuracy of 0.9910 and a multi-source balanced accuracy of 0.9677—is the direct outcome of our careful methodological choices.

The foundation of this success lies in the rigorous preprocessing pipeline (Section 3.2). The comparative evaluation confirmed that our hybrid filter effectively reduces noise while preserving critical features, validated by excellent PSNR (39.88 dB) and SSIM (0.8845) scores. This ensured the model was trained on high-quality data, which is crucial for achieving reliable performance.

Furthermore, the lightweight model architecture (Section 3.5) proved highly effective. By strategically pruning the VGG16 model and employing a dual-phase fine-tuning strategy, we achieved state-of-the-art accuracy while dramatically reducing computational cost. This result demonstrates that architectural efficiency and high performance are not mutually exclusive.

Finally, the most critical finding is the model's proven generalization capability, substantiated by our two-layered validation framework (Section 3.4). The model's stability was confirmed internally with K-Fold cross-validation, but more importantly, its real-world applicability was proven by successfully retraining and evaluating it on a large, heterogeneous multi-source dataset (as detailed in Section 4.4). This comprehensive validation, supported by a full suite of metrics (Section 3.7), addresses a critical gap in the literature and confirms that the model is ready for clinical deployment.

Comparison with Existing Methods:

As evidenced by the comparative summary in Table 15, our proposed framework addresses several critical gaps in the existing literature. While numerous studies have achieved high accuracy, our work distinguishes itself through a combination of methodological rigor, proven generalization, and computational efficiency.

First and most critically, our study confronts the "generalization gap" head-on. The table clearly shows that the vast majority of prior works, including those with near-perfect accuracy scores like Tandon *et al.* (2022), Ghosh *et al.* (2023), and Jassim *et al.* (2024), limit their validation to a single, often small, dataset. In contrast, our framework not only achieves a competitive accuracy of 99.1% on its primary single-source dataset but also demonstrates robust performance on a large, challenging multi-source dataset with a balanced accuracy of 96.9%. This dual-level validation provides a much higher

degree of confidence in the model's real-world clinical applicability; a feature largely is absent in the compared literature. Second, our approach balances high performance with computational efficiency. Many state-of-the-art results in the table are achieved using computationally expensive ensembles (Jassim *et al*., 2024; Kumaran *et al*., 2024) or large custom models (Gupta *et al*., 2023). Our work takes a more practical approach by using structured pruning to create a lightweight yet powerful model. This makes our solution more feasible for deployment in clinical settings with limited resources, a crucial consideration that is often overlooked in the pursuit of marginal accuracy gains.

Finally, our framework emphasizes methodological transparency. We combine advanced preprocessing with quantitative validation to ensure data quality. Also, we integrate Grad-CAM for interpretability, aligning our work with the best practices seen in the most recent studies by Kumaran *et al.* (2024) and Klangbunrueang *et al.* (2025). In summary, while prior works excel in specific areas, our study presents a more holistic and clinically viable solution by delivering a model that is simultaneously accurate, efficient, interpretable, and—most importantly—proven to generalize across diverse data sources.

Table 15: Comparison of Methodologies and Performance in Recent Lung Cancer Classification Literature
This table provides a structured comparison of recent studies. The final row highlights our proposed model to facilitate a direct comparison. Note: Acc: Accuracy, Sens: Sensitivity, Spec: Specificity, Prec.: Precision, Rec: Recall, MCC: Matthews Correlation Coefficient, AUC: Area Under the Curve, XAI: Explainable AI, --: Not specified or Not Applicable.

Study	CT Lung Dataset & Size	Preprocessing	Augmentation/ Data Balancing	Classification Type	XAI	Model & Results
Anand *et al.* (2022)	IQ-OTH/NCCD (977 images)	Basic (DICOM to JPG, Resize, Normalize)	Basic (flip, rotate, zoom, Brightness)	Binary	--	• VGG16: Acc. 0.96, Sens. 0.94, Spec. 0.96
Tandon *et al.* (2022)	LIDC-IDRI (7,500 images)	Basic (Resize, Normalize)	Flip, Rotate	Binary	--	• VCNet (Hybrid): Acc. 0.99, F1. 0.991, AUC. 0.991
Naseer *et al.*, (2023)	LUNA16 (888 CT scans)	Resize, Normalize, Patch Extraction, Segmentation	Patch-based	Binary	--	• Modified AlexNet-SVM: Acc. 0.979, Sens. 0.988, F1. 0.977
Sangeetha *et al.* (2023)	Kaggle Hist. & CT (750 images)	Denoising, Normalize, CLAHE	Basic Augmentation	Multi-class	--	• ResNet50 Histopathology Acc: 0.988 CT scan Acc: 0.847
Gupta *et al.* (2023)	SPIE-AAPM (18,000 images divided into 4 small sub-datasets for training: 2805, 2600, 2431, 3561)	-Resize, Normalize -Denoising (Gaussian and median filters) -CLAHE)	Shearing, Zooming, Horizontal Flipping, Fill mode.	Multi-class	--	• Custom CNN: Acc: ranging from 0.952 to 0.998 across different small sub datasets, Avg F1. 0.97, Avg Rec. 0.977
Ghosh *et al.* (2023)	IQ-OTH/NCCD (1,097 images)	-Resize, Smoothing, Threshold, Histogram, Equalization	Flip, Rotate	Multi-class	--	• CNN: Acc. 0.867, Prec. 0.939, Rec.0.704, AUC 0.946, F1: 0.74 • VGG16: Acc. 0.982, Prec. 0.956, Rec. 1.0, AUC:1.0, F1: 0.979
Al-Shouka and Alheeti (2023)	Kaggle Chest CT (1,200 images)	Resize, Normalize	Rotation, Shift, Shear, Zoom, flip, Fill mode.	Binary	--	• ResNet: Acc 0.90, loss 0.16. • MobileNetV2: Acc. 0.93, loss, 0.16. • Xception: Acc. 0.92, 0.19 loss. • VGG16: Acc. 0.91, 0.18 loss.
Gugulothu and Balaji (2023)	LIDC-IDRI (subsets)	Denoising, Contrast Enhancement	--	Binary	--	• HDE-NN (Hybrid): Acc. 0.963, Sens. 0.952
Benamara *et al*. (2024)	IQ-OTH/NCCD (1,097 images)	Resizing, Brightness, Sharpness	--	Multi-class	--	• DenseNet169 (Modified): Acc.1.0

Kumaran *et al.* (2024)	IQ-OTH/NCCD (1,097 images)	Resize, RGB convert, Normalize	SMOTE, Class Weighting	Multi-class	Grad-CAM	• Ensemble (VGG16 + ResNet50 + InceptionV3): Acc. 0.981, Bal. Acc. 0.969, MCC. 0.968, Kappa: 0.969, MSE. 0.061, RMSE. 0.249, MAE. 0.033
Alheeti *et al.* (2024).	Kaggle Chest CT-(1,000 images)	Brightness, Sharpness Conversion, Resizing.	Flip, Rotate	Binary	--	• MobileNetV2: Acc. 0.98, F1. 0.98
Jassim *et al.* (2024).	Kaggle Chest CT (1,000 images)	Resize, RGB convert	flipping, rotation, scaling	Multi-class	--	•Ensemble (ResNet50/101+EfficientNetB 3) • Validation Acc. 0.994
Klangbunrueang *et al.* (2025).	IQ-OTH/NCCD (1,097)	Image resizing, normalization	Rotation, Scaling, Flipping	Multi-class	Grad-CAM	• VGG16: Acc. 0.981 • MobileNetV2: Acc. 0.971 • ResNet50: Acc. 0.933
This Study	**• IQ-OTH/NCCD (1,097 Single-source) • Multi-sources (29,546 images)**	**Hybrid Filter + CLAHE (Quantitatively Validated)**	**SMOTE+ Hybrid Augmentation**	**Multi-class**	**Grad - CAM**	**Modified VGG16: - (Single Source Dataset)** Acc. 0.991, Bal. Acc. 0.9722, MCC. 0.984, Kappa &QWK: 0.984, MSE. 0.036, F1. 0.9911, AUC. 0.989 **- (Multi-Sources Dataset)** Acc. 0.964, Bal. Acc. 0.969, MCC. 0.943, Kappa: 0.942, MSE. 0.0193, F1. 0.966, QWK. 0.949 AUC. 0.998

Study's Limitations and Future Directions:

Despite the promising results, this study has several limitations that open avenues for future research. First, our study focused on enhancing the VGG16 architecture. While this demonstrated the power of our methodology, future work should extend this framework to other architectures, such as EfficientNet or vision transformers, to explore potential performance trade-offs.

Second, while our model accurately classifies entire images, it does not perform lesion segmentation. Integrating an automated segmentation module is a critical next step that would enhance the system's clinical utility by providing precise lesion localization and boundaries. Finally, while Grad-CAM improved model interpretability, clinical expert validation is needed. Future work can include deploying the model as an app/API for real-time clinical use.

CONCLUSION

This study successfully developed and validated a lightweight, accurate, and robust deep learning framework for automated lung cancer classification from CT images. By systematically addressing common challenges in data quality, model efficiency, and validation, we have created a solution poised for real-world clinical application. Our approach integrated a quantitatively validated preprocessing pipeline, a streamlined VGG16 architecture improved through pruning and fine-tuning, and a rigorous multi-layered validation strategy. The resulting model demonstrated outstanding and highly generalizable performance on both single-source and challenging multi-source datasets. Furthermore, the inclusion of Grad-CAM for model interpretability enhances its clinical utility by providing transparent, visual evidence for its predictions. Ultimately, this work presents a holistic and methodologically sound framework that sets a high standard for developing clinically viable computer-aided diagnosis (CAD) systems for the early and reliable detection of lung cancer.

Acknowledgments:

Not applicable.

Ethical Statement:

This study is a secondary analysis of publicly available, fully anonymized datasets. All data were originally collected under protocols approved by the respective institutional review boards (IRBs) of the data providers. As the research involved no direct interaction with human participants and exclusively utilized de-identified, pre-existing data, additional ethical approval was not required for this secondary analysis.

Author Contributions:

All authors have read and approved the final manuscript and agree to be accountable for all aspects of the work. **M. S. R.** Conceptualization, methodology, data acquisition, analysis, software development, validation, visualization, and original draft preparation. **M. A. S.,** Conceptualization, supervision, manuscript review, and critical revision for intellectual content.

Declaration of Competing Interests:

The author(s) declare that they have no competing interests.

Funding:

The author(s) received no specific funding for this work.

REFERENCES

Al Najjar, Y. (2024). Comparative analysis of image quality assessment metrics: MSE, PSNR, SSIM, and FSIM. International *Journal of Science and Research*, 13(3), 110–114. https://doi.org/10.21275/SR24302013533

Alheeti, K. M. A., Al-Shouka, T. T., Majeed, S. H., & Ahmed, A. A. (2024). Lung cancer detection using machine learning and deep learning models. In *2024 21st International Multi-Conference on Systems, Signals & Devices (SSD)* (pp. 63–69). IEEE. https://doi.org/10.1109/SSD61670.2024.1054950.

Al-Shouka, T. T., & Alheeti, K. M. A. (2023). A transfer learning for intelligent prediction of lung cancer detection. In *2023 Al-Sadiq International Conference on Communication and Information Technology (AICCIT)* (pp. 54–59). IEEE. https://doi.org/10.1109/AICCIT57614.2023.10217967.

Al-Yasriy, H. F., Al-Husieny, M. S., Mohsen, F. Y., Khalil, E. A., & Hassan, Z. S. (2020). Diagnosis of lung cancer based on CT scans using CNN. *IOP Conference Series: Materials Science and Engineering, 928*, 032033. https://www.kaggle.com/datasets/hamdallak/the-iqothnccd-lung-cancer-dataset/data.

Anand, R., Rao, N., & Sumukh, D. (2022). Lung cancer detection and prediction using deep learning. *International Journal of Engineering Applied Sciences and Technology*.

Anusha, M., & Reddy, D. S. (2023). Lung carcinoma diagnosis and classification using deep learning. *2023 4th International Conference for Emerging Technology (INCET)*, 1–4. IEEE. https://doi.org/10.1109/INCET57972.2023.10170615.

Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, L., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G. S., Naidich, D. P., & Shetty, S. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature Medicine, 25(6), 954–961. https://doi.org/10.1038/s41591-019-0447-x.

Benamara, Z., Zehani, S., & Zitouni, A. (2024). The effect of fully connected layers in different pre-trained CNN architectures on the enhancement of lung cancer classification. In *2024 8th International Conference on Image and Signal Processing and their Applications (ISPA)* (pp. 1–6). IEEE. https://doi.org/10.1109/ISPA59904.2024.10536830.

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer *Journal for Clinicians*, 74(3), 229–263. https://doi.org/10.3322/caac.21834.

Chang, Y., Jung, C., Ke, P., Song, H., & Hwang, J. (2018). Automatic contrast-limited adaptive histogram equalization with dual gamma correction. *IEEE Access*, *6*, 11782–11792. https://doi.org/10.1109/ACCESS.2018.2797872

Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2018, March). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 839–847). IEEE. https://doi.org/10.1109/WACV.2018.00097.

Gupta, D., Dawn, S., & others. (2023). *Detection and staging of lung cancer from CT scan images by deep learning*. In 2023 International Conference on Disruptive Technologies (ICDT) (pp. 274–278). IEEE. https://doi.org/10.1109/ICDT57929.2023.10151194.

Ghosh, R., Ahamed, A., Sadhukhan, B., & Das, N. (2023). Lung nodule classification using MobileNet transfer learning. In *2023 9th International Conference on Smart Computing and Communications (ICSCC)* (pp. 290–295). IEEE. https://doi.org/10.1109/ICSCC59169.2023.10335043.

Gugulothu, V. K., & Balaji, S. (2024). *RETRACTED ARTICLE: An early prediction and classification of lung nodule diagnosis on CT images based on hybrid deep learning techniques*. *Multimedia Tools and Applications, 83*, 1041–1061. https://doi.org/10.1007/s11042-023-15802-2

Ibrahim, W. R., & Mahmood, M. R. (2023). Classified Covid-19 By Densenet121-Based Deep Transfer Learning From Ct-Scan Images. *Science Journal of University of Zakho*, *11*(4), 571 – https://doi.org/10.25271/sjuoz.2023.11.4.1166.

Jassim, O. A., Abed, M. J., & Saied, Z. H. (2024). Deep learning techniques in the cancer-related medical domain: A transfer deep learning ensemble model for lung cancer prediction. *Baghdad Science Journal, 21*(3). https://doi.org/10.21123/bsj.2023.8340.

Klangbunrueang, R., Pookduang, P., Chansanam, W., & Lunrasri, T. (2025, February). AI-powered lung cancer detection: Assessing VGG16 and CNN architectures for CT scan image classification. *Informatics, 12*(1), 18. MDPI. https://doi.org/10.3390/informatics12010018.

Kumaran, Y. S., Jeya, J. J., T. R, M., Khan, S. B., Alzahrani, S., & Alojail, M. (2024). Explainable lung cancer classification with ensemble transfer learning of VGG16, ResNet50 and InceptionV3 using Grad-CAM. *BMC Medical Imaging, 24*(1), 176. https://doi.org/10.1186/s12880-024-01345-x.

Leiter, A., Veluswamy, R. R., & Wisnivesky, J. P. (2023). The global burden of lung cancer: Current status and future trends. Nature Reviews Clinical Oncology, 20, 624–639. https://doi.org/10.1038/s41571-023-00798-3.

Murad, S. H., Awlla, A. H., & Moahmmed, B. T. (2023). Prediction lung cancer based critical factors using machine learning. *Science Journal of University of*

Zakho, 11(3), 447–452. https://doi.org/10.25271/sjuoz.2023.11.3.1105

National Lung Screening Trial Research Team (NLST). (2011). Reduced lung-cancer mortality with low-dose computed tomographic screening. New England *Journal of Medicine*, 365(5), 395–409. https://doi.org/10.1056/NEJMoa1102873.

Naseer, I., Akram, S., Masood, T., Rashid, M., & Jaffar, A. (2023). Lung cancer classification using modified U-Net based lobe segmentation and nodule detection. *IEEE Access, 11*, 60279–60291. https://doi.org/10.1109/ACCESS.2023.3285821.

Park, S., Park, H., Lee, S. M., Kim, H., & Goo, J. M. (2022). Application of computer-aided diagnosis for Lung-RADS categorization in CT screening for lung cancer: Effect on inter-reader agreement. *European Radiology, 32*(2), 1054–1064. https://doi.org/10.1007/s00330-021-08202-3

Ravindra, C., Nalband, A. H., Kumar, G., Basheer, S., & Ravindra, M. (2024). From pixels to prognosis: A deep dive into lung cancer subtype classification using transfer learning. In *2024 IEEE International Conference on Contemporary Computing and Communications (InC4)* (Vol. 1, pp. 1–6). IEEE. https://doi.org/10.1109/InC460750.2024.10649168.

Rodrigues, R., Lévêque, L., Gutiérrez, J., Jebbari, H., Outtas, M., Zhang, L., Chetouani, A., Al-Juboori, S., Martini, M. G., & Pinheiro, A. M. G. (2024). Objective quality assessment of medical images and videos: Review and challenges. Multimedia Tools and Applications, 1–34. https://doi.org/10.1007/s11042-024-20292-x.

Sangeetha, M., Devi, R. M., Gunasekaran, H., Venkatesan, R., Ramalakshmi, K., & Murugesan, P. (2023). Deep residual learning for lung cancer nodules detection and classification. In *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 907–912). IEEE. https://doi.org/10.1109/ICCMC56507.2023.10083783.

Singh, A., & Kamath, S. (2024). Preprocessing of CT scans for lung cancer detection. In *2024 4th International Conference on Intelligent Technologies (CONIT)* (pp. 1–4). IEEE. https://doi.org/10.1109/CONIT61985.2024.10626594.

Tandon, R., Agrawal, S., Chang, A., & Band, S. S. (2022). VCNet: Hybrid deep learning model for detection and classification of lung carcinoma using chest radiographs. *Frontiers in Public Health, 10*, 894920. https://doi.org/10.3389/fpubh.2022.894920.

World Health Organization (WHO). (2023). *Lung cancer* (Report No. WHO/2023/LC_FS). https://www.who.int/news-room/fact-sheets/detail/lung-cancer.