



RESEARCH ARTICLE

Deep Convolutional Neural Network–Based Multiclass Classification of Pulmonary Lesions Using Computed Tomography Imagery from the IQ-OTH/NCCD Dataset

Dharm Patel, Hetkumar Patel, Wisam Bukaita* 

*Math and Computer Science

Department, Lawrence Technological University, Southfield, USA

<https://orcid.org/0000-0001-6255-3848> (Wisam Bukaita),

<https://orcid.org/0009-0005-1551-6055> (Hetskumar Patel)



OPEN ACCESS

PUBLISHED

30 April 2026

CITATION

Patel, D., et al., 2026. Deep Convolutional Neural Network–Based Multiclass Classification of Pulmonary Lesions Using Computed Tomography Imagery from the IQ-OTH/NCCD Dataset. *Medical Research Archives*, [online] 14(4).

COPYRIGHT

© 2026 European Society of Medicine. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

ISSN

2375-1924

ABSTRACT

This study presents an advanced deep learning framework for automated multiclass detection and classification of pulmonary lesions—normal, benign, and malignant—using Computed Tomography imagery from the publicly available the Iraq Oncology Teaching Hospital/National Center for Cancer Diseases chest computed tomography dataset dataset. The proposed model employs a 16-layer convolutional neural network developed by the Visual Geometry Group Network as the primary feature extraction backbone, enhanced through customized preprocessing operations and data augmentation strategies designed to improve model generalization. The dataset consists of 1,190 de-identified Computed Tomography scan images, enabling the model to autonomously learn discriminative radiological features and perform diagnostic classification with reduced dependence on subjective human interpretation. The training pipeline integrates transfer learning, intensity normalization, and class-balanced sampling to mitigate dataset imbalance and strengthen model robustness. Experimental evaluation yielded strong performance outcomes, including an overall classification accuracy of 97.73%, class-specific precision scores of 100% (Benign), 99% (Malignant), and 95% (Normal), and an Area Under the Curve of 99.34%. A confusion matrix analysis further validated the model’s reliability, particularly in the accurate discrimination of malignant lesions from benign and healthy tissue. Comparative analyses against traditional machine learning classifiers demonstrated the superior effectiveness of the 16-layer convolutional neural network developed by the Visual Geometry Group Network based deep transfer learning architecture. The developed framework offers a scalable, cost-efficient, and automated diagnostic support tool for early lung cancer detection. Furthermore, its interpretability is strengthened through the introduction of two novel metrics the Multistage Diagnostic Confidence Index and the Patient Stability Index which provide enhanced transparency in model decision-making. These findings highlight the framework’s substantial potential for integration into clinical decision-support systems and early screening workflows.

Keywords: Deep Convolutional Networks, Pulmonary Nodule Classification, Medical Image Analysis, Transfer Learning, Computer-Aided Diagnosis

1. Introduction

Lung cancer remains one of the most formidable public health challenges globally, standing as a leading cause of cancer-related mortality. The prognosis for patients is intrinsically linked to the stage of detection, emphasizing the critical need for advanced, accurate, and scalable screening modalities. Computed Tomography (CT) imaging has long served as the clinical standard for identifying and characterizing pulmonary lesions, offering detailed spatial resolution essential for distinguishing between normal tissue, benign growths, and early-stage malignancies.

Despite the inherent diagnostic value of CT scans, traditional radiological interpretation presents several fundamental limitations: it is inherently time-consuming, highly dependent on the subjective expertise of the interpreting radiologist, and constrained by inter-rater variability, all of which hinder early-stage screening efforts in high-volume clinical settings. The advancement of deep learning (DL) and Artificial Intelligence (AI) in healthcare offers a powerful paradigm shift, enabling automated feature extraction and pattern recognition directly from complex medical imagery.

While numerous deep learning models have been proposed for pulmonary analysis, existing research often concentrates on binary classification tasks (cancer vs. non-cancer) or relies on datasets with limited diversity. This results in frameworks that often lack the necessary reliability and interpretability for real-world clinical deployment, particularly in distinguishing the clinically critical third class—benign lesions—from both normal and malignant findings.

This study addresses these deficits by introducing a robust, interpretable, and multiclass deep learning framework for the classification of lung cancer using the IQ-OTH/NCCD CT scan dataset. The primary contributions of this work are threefold:

- 1) **Multiclass Classification:** Development of a high-performance model capable of accurately classifying CT slices into three distinct

categories: normal, benign, and malignant.

- 2) **Transfer Learning Optimization:** Leveraging the VGG16 architecture, optimized through targeted preprocessing, data augmentation, and class-balanced sampling to achieve exceptional generalization capability on the imbalanced medical dataset.
- 3) **Enhanced Interpretability and Reliability:** Integration of novel, clinically relevant metrics the Multistage Diagnostic Confidence Index (MDCI) and the Patient Stability Index (PSI) to quantitatively evaluate the diagnostic trust and prediction consistency, moving beyond conventional performance scores.

2. Theoretical Background and Prior Studies

Artificial intelligence (AI) has rapidly transformed medical imaging by enabling automated, scalable, and objective analysis of complex radiological data. In oncology, where early and accurate diagnosis directly influences patient outcomes, AI-driven systems have shown significant potential to enhance diagnostic efficiency, reduce inter-observer variability, and support clinical decision-making. The integration of deep learning (DL) with medical imaging has shifted traditional workflows away from purely manual interpretation toward data-driven, reproducible diagnostic frameworks^{13, 15}.

Recent advances in healthcare-oriented deep learning further demonstrate the broad clinical applicability of modern neural architectures across diverse disease domains. For example, deep vision models have been successfully applied to brain tumor detection with real-time MRI analysis²⁹, automated Alzheimer's disease diagnosis and localization³⁰, and prediction of Type 1 diabetes progression from continuous glucose monitoring data³¹, highlighting the scalability of DL-based solutions beyond single-organ imaging tasks. Moreover, machine learning–driven cardiovascular risk prediction models⁴ reinforce the expanding role of intelligent systems in clinical decision support across medical specialties.

2.1. CLINICAL FOUNDATIONS OF PULMONARY NODULE ASSESSMENT

Computed Tomography (CT) imaging remains the gold standard for pulmonary nodule detection and characterization due to its superior spatial resolution and sensitivity. Large clinical studies and guidelines have established CT as the primary modality for lung cancer screening and nodule management. Foundational resources such as the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI)¹ have played a critical role in standardizing annotated CT datasets, enabling reproducible research and algorithm benchmarking.

In clinical practice, CT is often complemented by positron emission tomography combined with CT (PET/CT), which provides functional metabolic information that can improve malignancy assessment. Chung et al.¹⁶ review the complementary roles of CT and PET/CT in pulmonary nodule evaluation, showing that CT offers detailed structural characterization, while

PET/CT adds metabolic insight to enhance malignancy detection, with recognized limitations for small nodules and inflammatory lesions.

Clinical practice guidelines from the Fleischner Society², the American College of Chest Physicians (ACCP)³, and the British Thoracic Society⁸ emphasize the importance of accurate risk stratification of pulmonary nodules to guide follow-up and intervention. Studies have shown that CT-based screening significantly improves survival rates for early-stage lung cancer when malignancies are detected promptly^{6, 23}. However, conventional radiological assessment remains time-consuming and subject to inter-rater variability, particularly when distinguishing benign from malignant nodules^{5, 7, 17, 27}.

2.2. EARLY COMPUTER-AIDED DIAGNOSIS AND RADIOMICS APPROACHES

Before the rise of deep learning, computer-aided diagnosis (CAD) systems relied on handcrafted features and statistical models to estimate malignancy risk. Techniques such as CT texture analysis²² and

radiomics-based feature extraction²⁵ demonstrated promising results in differentiating benign and malignant lesions. Probabilistic malignancy models further advanced risk estimation by integrating clinical and imaging features^{7, 9}. While effective, these approaches depend heavily on feature engineering and often struggle to generalize across datasets and imaging protocols.

2.3. EMERGENCE OF DEEP LEARNING IN PULMONARY IMAGING

The introduction of deep Convolutional Neural Networks (CNNs) marked a paradigm shift in medical image analysis. Following the success of CNNs in large-scale visual recognition tasks¹⁹, researchers adapted deep architectures to CT-based pulmonary nodule detection and classification. CNN-based systems demonstrated strong capability in extracting hierarchical and discriminative features directly from raw image data, eliminating the need for handcrafted descriptors^{10, 11, 14}. The adoption of Rectified Linear Units (ReLU), introduced by Nair and Hinton¹⁸, further improved deep network training by reducing vanishing gradient effects and accelerating convergence, contributing to the practical success of modern CNN architectures in medical imaging.

Several studies have reported high accuracy in pulmonary nodule detection and classification using CNNs. Setio et al.¹¹ validated and compared multiple automated detection algorithms, highlighting the effectiveness of deep learning–based approaches. Huang et al.¹² and Li et al.¹⁴ further demonstrated that CNN-based models achieve competitive performance in both detection and malignancy classification tasks. Lightweight CNN architectures have also shown promise in balancing accuracy and computational efficiency, making them suitable for real-time and resource-constrained environments¹⁵.

2.4. TRANSFER LEARNING AND ADVANCED DEEP ARCHITECTURES

Transfer learning has emerged as a critical strategy for addressing limited annotated medical datasets. Pretrained CNNs such as VGG16, ResNet, and

Inception architectures leverage knowledge learned from large-scale natural image datasets to improve generalization in medical imaging tasks¹³. Studies have consistently shown that transfer learning significantly enhances performance in pulmonary nodule classification^{12, 14}.

More recent work has explored object detection frameworks and attention-based models. Xu et al.²⁰ applied YOLO-based architectures for automated lung nodule detection, achieving strong detection accuracy. Transformer-based models and hybrid CNN–Transformer architectures have further advanced the field by capturing long-range contextual dependencies that traditional CNNs may overlook²⁶. These approaches improve global feature modeling and offer enhanced interpretability through attention mechanisms.

2.5. EXPLAINABILITY, RELIABILITY, AND CLINICAL TRUST

Despite impressive performance gains, the lack of interpretability remains a major barrier to clinical adoption of deep learning models. Explainable AI (XAI) techniques such as Grad-CAM and LIME have been introduced to visualize model attention and ensure that predictions are based on clinically relevant features rather than spurious artifacts^{13, 21}. Systematic reviews highlight that without explainability and reliability assessment, high-performing AI systems may still fail in real-world clinical settings²¹.

Moreover, most existing studies rely predominantly on conventional performance metrics such as accuracy, F1-score, and AUC. While informative, these metrics do not quantify diagnostic confidence, slice-level consistency, or patient-level stability factors that are critical for clinical trust and longitudinal assessment across CT volumes.

2.6. IDENTIFIED RESEARCH GAPS

A synthesis of the literature reveals three persistent limitations:

- 1) Limited Multiclass Focus: The majority of studies emphasize binary classification (benign vs. malignant), neglecting the clinically essential

distinction among normal, benign, and malignant categories required for effective triage and follow-up planning^{7, 9, 24}.

- 2) Lack of Reliability and Stability Metrics: Existing evaluations rarely assess prediction consistency across multiple CT slices or quantify diagnostic confidence, limiting clinical interpretability and trustworthiness²¹.
- 3) Data Imbalance and Generalization Challenges: Medical datasets are inherently imbalanced, often leading to biased models that favor majority classes. While augmentation and resampling are commonly applied, more robust strategies are required to ensure reliable minority-class performance^{15, 25}.

2.7. MOTIVATION FOR THE PRESENT STUDY

The present study addresses these gaps by developing a robust multiclass pulmonary lesion classification framework using the IQ-OTH/NCCD CT dataset. An optimized VGG16-based transfer learning architecture is employed to balance performance and computational efficiency. To enhance clinical relevance, two novel reliability metrics the Multistage Diagnostic Confidence Index (MDCI) and the Patient Stability Index (PSI) are introduced to evaluate diagnostic confidence and prediction consistency across CT slices.

3. Methodology

This study introduces a robust, interpretable deep transfer learning framework for the multiclass classification of pulmonary lesions using Computed Tomography (CT) imagery from the IQ-OTH/NCCD dataset. The methodology integrates specialized medical image preprocessing, feature extraction via an optimized VGG16 model, and a comprehensive evaluation using both conventional performance metrics and novel clinical reliability indices.

3.1. DATA SOURCE AND STRUCTURE

The foundation of this study is the IQ-OTH/NCCD Lung Cancer Dataset, a publicly accessible repository of de-identified Computed Tomography (CT) imagery. This dataset was selected to facilitate the

development and evaluation of a robust multiclass classification model for pulmonary lesions. The dataset is comprised of 1,190 CT scan slices sourced from 110 distinct patient cases. These cases are categorically partitioned into three clinically significant diagnostic classes, reflecting the intrinsic challenge of class imbalance commonly observed in medical datasets:

- 1) Normal: 55 cases
- 2) Benign: 15 cases
- 3) Malignant: 40 cases

The original CT scans were acquired using a Siemens SOMATOM scanner, adhering to a standardized imaging protocol that specified a tube voltage of 120 kV, a slice thickness of 1 mm, a window width ranging from 350–1200 HU (Hounsfield Units), and a window center set between 50 – 600 HU. Prior to ingestion by the deep learning architecture, all digital imaging and communications in medicine (DICOM) files were converted to the JPEG format to ensure compatibility with the computational processing pipeline.

3.2. DATA PREPROCESSING AND AUGMENTATION

To ensure diagnostic quality, uniformity, and mitigate the small size and imbalance of the medical dataset, extensive preprocessing and augmentation were performed:

- 1) Image Standardization:
 - a) Resizing: All images were resized to 224 pixels to standardize the input dimension for the VGG16 model.
 - b) Normalization: Pixel intensity values were normalized to the [0, 1] range to improve training stability and convergence speed.
 - c) Color Conversion: Grayscale images were converted to RGB format to align with the VGG16 architecture's input requirements.
- 2) Image Enhancement:
 - a) Contrast Enhancement: Applied histogram equalization to highlight fine lung structures and lesions.

- b) Noise Reduction: Employed Gaussian filtering to reduce noise and artifacts inherent in CT scans.

- 3) Data Augmentation: Techniques (including rotation, flipping, and scaling) were applied to artificially expand the training dataset diversity, mitigate overfitting, and partially address the class imbalance issue.
- 4) Label Encoding: Images were encoded into three categorical classes: Normal, Benign, and Malignant.

3.3. DEEP TRANSFER LEARNING MODEL ARCHITECTURE

The proposed classification model is a hybrid deep transfer learning framework built upon the VGG16 architecture.

A. Feature Extraction Backbone

- 1) Model: A VGG16 model, pretrained on the ImageNet dataset, was utilized as the feature extraction backbone.
- 2) Transfer Learning: The convolutional layers of VGG16 were fine-tuned for the medical imaging context.
- 3) Classification Layer: The original final dense layers of VGG16 were replaced with a custom network tailored for the multiclass task:
 - a) A Fully Connected Dense Network was added to process the extracted feature vectors.
 - b) The final output layer is a three-node Softmax layer corresponding to the Normal, Benign, and Malignant classes.

B. Regularization and Optimization

To prevent overfitting and enhance convergence stability, the following modules were integrated:

- 4) Dropout: A Dropout rate of 0.4 was applied to the dense layers.
- 5) Batch Normalization: Used to normalize the activations of the hidden layers and stabilize the training process.

3.4. MODEL TRAINING

The dataset was split using a systematic partition ratio: 70% for training, 20% for validation, and 10% for testing.

- 1) Cross-Validation: Training employed 5-fold cross-validation to ensure every data subset was tested and to reduce the risk of subset-specific overfitting.
- 2) Hyperparameters:
 - a) Optimizer: Adam
 - b) Loss Function: Categorical Cross-Entropy
 - c) Batch Size: 32
 - d) Epochs: 10
 - e) Framework: TensorFlow/Keras (Python 3.10)
 - f) Hardware: Google Colab GPU (NVIDIA Tesla T4, 16 GB RAM)

3.5. EVALUATION AND RELIABILITY METRICS

Model performance was rigorously evaluated using a combination of traditional classification metrics and two novel, clinically-focused reliability indices.

- 1) Traditional Classification Metrics
- 2) Accuracy: Overall classification correctness.
- 3) Precision, Recall, and F1-Score: Measured for general and class-specific effectiveness.
- 4) AUC-ROC (Area Under the Curve - Receiver Operating Characteristic): Used to measure the model's discriminative capability across all classes.

- 5) Confusion Matrix: Provided a visualization of the true-positive and false-positive distributions for granular error analysis.
- 6) Novel Reliability Indices
- 7) To enhance model interpretability and clinical relevance, the following metrics were introduced:
 - 8) Multistage Diagnostic Confidence Index (MDCI): Quantifies the combined impact of lesion localization quality and classification reliability, balancing the performance of the visual feature extraction (VGG16) with diagnostic confidence.
 - 9) Patient Stability Index (PSI): Measures the prediction consistency across multiple CT slices of the *same patient*, simulating real-world diagnostic review and reflecting stable diagnostic behavior despite variations in image presentation.

4. Model Architecture

The core of the proposed framework for multiclass pulmonary lesion classification is a deep transfer learning model built upon the VGG16 architecture. This architecture was chosen as the feature extraction backbone due to its proven capability in extracting rich, hierarchical visual features from complex imagery, which is essential for detailed analysis of Computed Tomography (CT) scans. The general structure of a deep convolutional network, comprising sequential convolutional blocks followed by fully connected layers, is conceptualized in Figure 1.

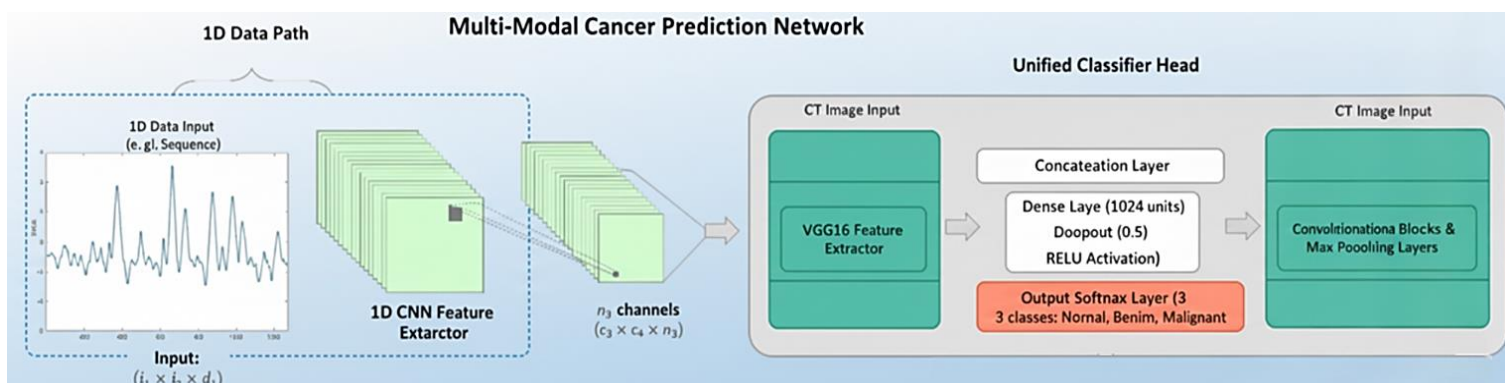


Figure 1. Hybrid VGG16-Based Deep Learning Model Architecture for Pulmonary Lesion Classification.

The proposed model applies a hybrid deep transfer learning strategy built on the VGG16 architecture to classify CT scan images into Normal, Benign, and Malignant categories. The design consists of two main components: a VGG16 feature extractor and a custom classification head.

4.1. FEATURE EXTRACTION USING VGG16

VGG16, pretrained on ImageNet, is used as the fixed convolutional backbone. Its stacked convolution–max-pooling layers extract increasingly complex spatial patterns from the 224×224 CT slices, capturing both low-level textures and high-level lesion characteristics. The final feature maps are flattened into a 25088-dimensional vector, which serves as the input to the classifier.

4.2. CUSTOM CLASSIFICATION HEAD

A new dense network replaces VGG16’s original fully connected layers to address the three-class medical diagnosis task.

- 1) Dense Layer 1: Reduces the 25088-dimensional feature vector to 1024 units, followed by a ReLU activation and Dropout (0.5) to limit overfitting.
- 2) Dense Layer 2: Further reduces the representation to 512 units, again with ReLU activation to enhance nonlinear learning.
- 3) Output Layer: A three-node Softmax layer produces the final probability scores corresponding to *Normal*, *Benign*, and *Malignant* lesion classes.

Batch Normalization and dropout integrated within the dense layers improve training stability and generalization.

Model Performance

This optimized hybrid VGG16 configuration demonstrated superior classification accuracy compared to deeper or more complex architectures such as ResNet50 and InceptionV3, while remaining computationally efficient for CT-based pulmonary lesion analysis.

5. Results

The empirical results of the proposed VGG16-based deep learning framework for multiclass pulmonary lesion classification draw upon the curated dataset, the defined training configuration, and the evaluation metrics outlined earlier. These findings illustrate key aspects of the model’s behavior across preprocessing, training, and validation, highlighting its performance and reliability throughout the classification pipeline.

5.1. DATASET CHARACTERISTICS AND PREPROCESSING OUTCOMES

The characteristics of the IQ-OTH/NCCD dataset and the outcomes of the preprocessing pipeline are shown in Figure 2, which depicts the distribution of the three diagnostic classes: Normal (46.9%), Malignant (36.3%), and Benign (16.8%). The evident class imbalance raises the risk of biasing the model toward the majority categories, necessitating the use of data augmentation and class-balanced sampling to improve generalization, particularly for the underrepresented benign class. Prior to model training, a comprehensive data-cleaning process was conducted to ensure the reliability of the inputs. This included the removal of corrupted CT slices, verification of class labels, and elimination of duplicate images. The subsequent preprocessing steps—grayscale-to-RGB conversion, resizing to 224 × 224 pixels, intensity normalization, and Gaussian noise reduction—produced standardized, high-quality images suitable for the VGG16 architecture. These procedures enhanced feature consistency, improved contrast, and mitigated scanner-related variability, thereby strengthening the foundation for effective model learning.

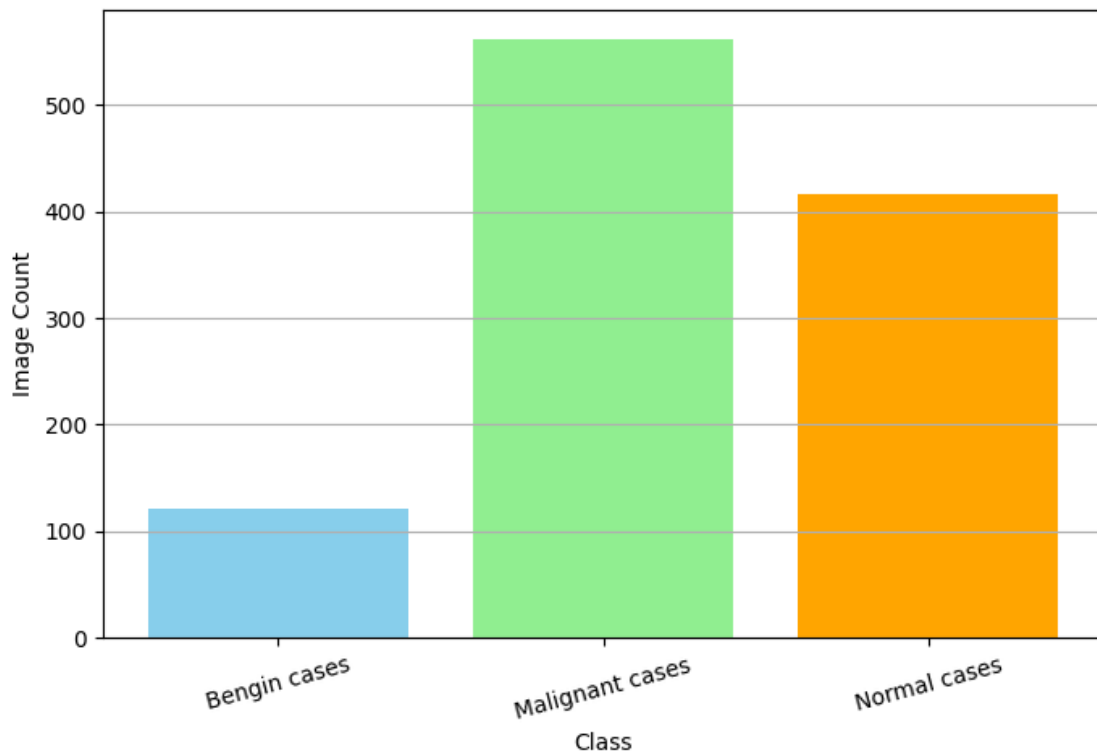


Figure 2. Class Distribution of Lung Cancer CT Images.

5.2. HANDLING CLASS IMBALANCE AND DATA AUGMENTATION

Class imbalance within the dataset was mitigated through a set of targeted data augmentation strategies, including rotations, horizontal and vertical flips, elastic deformations, and controlled scaling operations. As illustrated in Figure 3, these transformations generated diverse representations of lesion morphology, intensity variations, and spatial orientations, thereby increasing the effective dataset size and enhancing the robustness of the training process. The augmentation procedures were particularly beneficial for the underrepresented benign class, improving its visibility during training and reducing the model's tendency to favor majority classes. Subsequent ablation experiments demonstrated that the removal of augmentation resulted in a 3% decrease in F1-score, confirming its essential role in improving model performance and generalization.

The class imbalance within the dataset was addressed using a set of targeted data augmentation strategies. These strategies included rotations, horizontal and vertical flips, elastic deformations, and controlled scaling operations. As illustrated in Figure 3, these

transformations created diverse representations of lesion morphology, intensity variations, and spatial orientations. This process was key to increasing the effective dataset size and enhancing the robustness of the training.

The augmentation was especially beneficial for the underrepresented benign class, improving its visibility during training and reducing the model's tendency to favor the majority classes. Subsequent ablation experiments confirmed the crucial role of augmentation, showing that its removal resulted in a 3% decrease in the F1-score, which verifies its essential role in improving model performance and generalization.

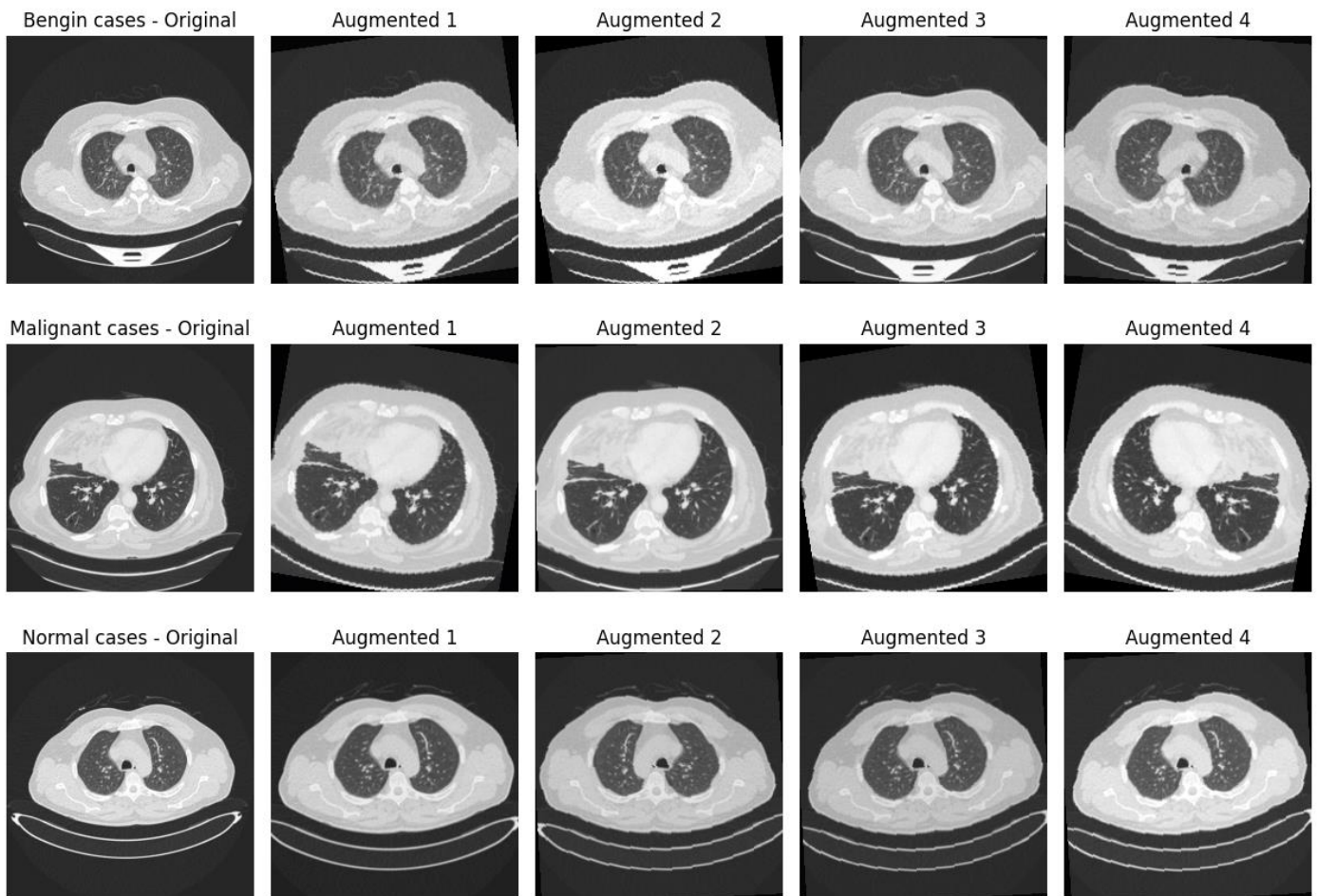


Figure 3. Data Augmentation Results Across Cognitive Conditions.

Figure 3 presents representative examples of the data augmentation process applied to each of the three classes, illustrating how the original dataset is systematically expanded and diversified. For instance, a single benign image yields four additional, distinct training samples, substantially increasing the number of examples available for underrepresented classes. This expansion plays a critical role in mitigating class imbalance and contributes to the development of a more robust and generalizable model.

The figure is organized into three rows corresponding to the Benign, Malignant, and Normal classes, and five columns representing the original CT slice followed by four augmented variants. The augmented images demonstrate the range of transformations applied during preprocessing. Rotational augmentations visible, for example, in the Malignant cases provide alternative spatial perspectives of the lesion, enabling the model to recognize diagnostic features independent of

orientation. Horizontal and vertical flips, clearly observed in the Benign cases, introduce mirrored representations that further enrich spatial diversity. Additionally, subtle variations in intensity and scale can be seen across all classes, reflecting modifications designed to improve invariance to scanner settings, illumination conditions, and acquisition differences. Together, these augmentation techniques enhance the variability of the training data and strengthen the model's capacity to generalize to unseen medical images.

5.3. MODEL TRAINING BEHAVIOR AND OPTIMIZATION SETUP

The model was trained using a 70–20–10 split for the training, validation, and testing sets, respectively. This partitioning strategy ensures that the majority of the data is used for parameter learning, while sufficient samples remain for unbiased performance evaluation and hyperparameter tuning. The network was optimized using the Adam optimizer, selected for its adaptive learning-rate mechanism,

which accelerates convergence and performs well in high-dimensional parameter spaces typical of deep convolutional networks. The categorical cross-entropy loss function was employed because it is the standard objective function for multiclass classification problems and directly penalizes deviations between predicted and true class probabilities.

A batch size of 32 was chosen as a balance between computational efficiency and gradient stability; smaller batches introduce excessive noise, whereas larger batches demand greater memory and may lead to suboptimal generalization. Training was conducted for 10 epochs, which empirical observations indicated were sufficient for convergence without overfitting, given the regularization mechanisms and data augmentation applied. The model was implemented in TensorFlow/Keras, a framework well-suited for rapid prototyping and GPU-accelerated training, and executed on an NVIDIA Tesla T4 GPU, whose architecture offers optimized performance for convolutional operations.

The 25,088-dimensional feature vectors extracted from the VGG16 convolutional backbone were subsequently forwarded through the custom fully connected classification head. This dense architecture allows the network to integrate global contextual information derived from the convolutional maps, facilitating the modeling of subtle differences across lesion types. The final Softmax layer then converts these learned representations into probabilistic outputs for the three diagnostic classes—Normal, Benign, and Malignant. This hierarchical feature-processing pipeline ensures that both localized lesion features and broader structural patterns are incorporated into the final classification decision, thereby improving diagnostic accuracy and model robustness.

5.4. CLASS-WISE PREDICTIVE BEHAVIOR AND COMPARATIVE ANALYSIS

The class-wise Precision–Recall (PR) analysis provides a detailed assessment of prediction confidence and reliability under the dataset’s inherent class

imbalance. Because PR curves directly measure precision as positive predictive value and recall for sensitivity, they offer a more reliable indicator of model performance in medical imaging tasks than accuracy-based metrics, which can be inflated by the abundance of true negatives.

The malignant class demonstrates the strongest predictive performance, characterized by consistently high precision and high recall across a broad range of decision thresholds. This behavior reflects the model’s ability to distinguish malignant lesions with strong confidence while minimizing false negatives, an essential property in clinical applications where missing a malignant case carries significant risk. These results suggest that the model successfully captures the distinctive morphological patterns associated with malignant tumors, supported by the deep hierarchical features extracted by the VGG16 backbone.

The benign class exhibits slightly lower but stable predictive performance. Its reduced curve elevation compared to the malignant class is expected given its limited representation in the dataset. However, the overall smoothness and upward trend of the benign PR profile indicate that augmentation strategies effectively increased feature variability and improved the model’s generalization capability for this minority class.

The normal class maintains high precision and moderate-to-high recall, consistent with the relatively homogeneous structure of non-pathological CT slices. This stability suggests that normal tissue characteristics are easier for the model to identify, contributing to consistently accurate classification.

Collectively, the PR behavior across all three classes confirms that the model sustains strong positive predictive value and sensitivity despite class imbalance and varying feature complexity among lesion types. This pattern of performance is clearly reflected in the class-wise PR curves, as shown in Figure 4, which visually demonstrate high discriminative ability and robust classification reliability across normal, benign, and malignant categories.

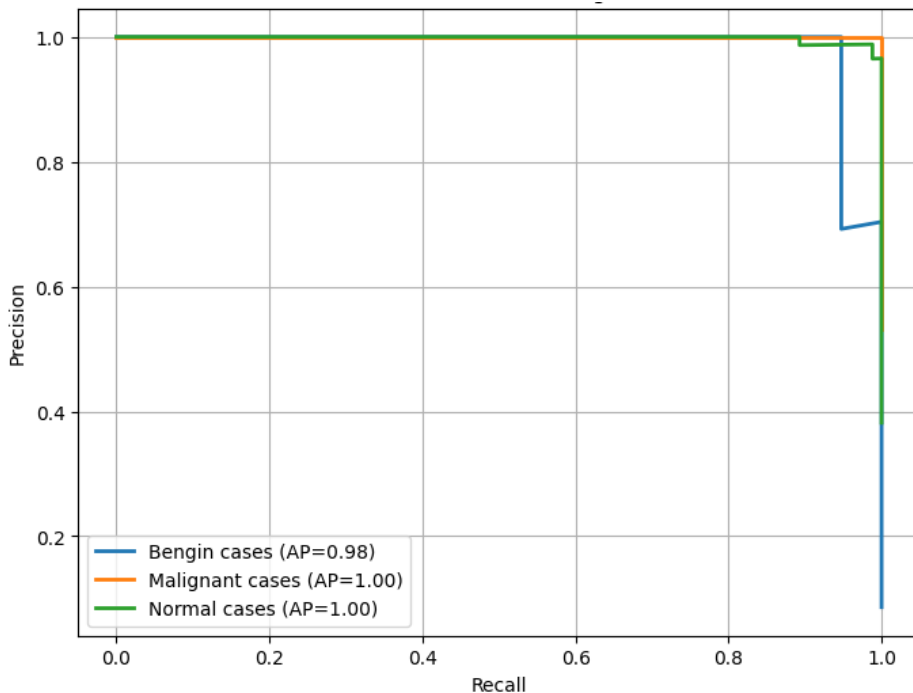


Figure 4. Precision-Recall Curves Across Lung Cancer Classes.

5.5. F1–CONFIDENCE BEHAVIOR

The F1-Confidence Curve shown in Figure 5 analyzes how the model's F1 Score changes as the Confidence Threshold varies. The F1 Score is a

measure of a model's accuracy, considering both precision and recall. The Confidence Threshold is the minimum probability a prediction must have to be considered positive for a specific class.

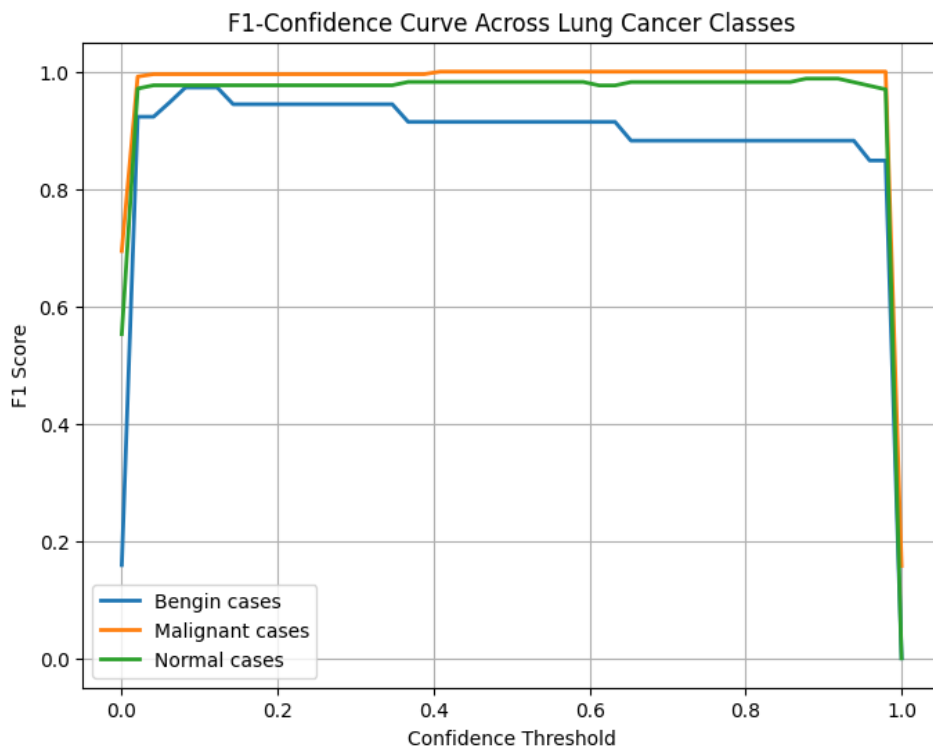


Figure 5. F1-Confidence Curve Across Lung Cancer Classes.

The curve for all three lung cancer classes (Benign cases, Malignant cases, and Normal cases) remains consistently high across a wide range of confidence thresholds approximately from 0.05 to 0.95. This consistent stability indicates several key strengths of the model:

- 1) Strong Feature Separability: The features the model learned are distinct enough that it can clearly separate the different lung cancer classes. This means the model isn't confused about which class an input belongs to.
- 2) Minimal Overconfident Misclassifications: The high F1 Score at high thresholds suggests that when the model is highly confident (assigns a high probability), its prediction is almost always correct, reducing the risk of confident errors.
- 3) Robust Calibration of Softmax Probabilities: The probability outputs (Softmax) from the model are well-calibrated. A well-calibrated model's predicted probability reflects the true likelihood of the prediction being correct. The F1 score stability confirms that the confidence scores are reliable indicators of correctness.

The stability of the F1 Score across different thresholds implies that lowering or raising the decision threshold does not significantly degrade the model's classification consistency. This robustness is highly desirable in clinical applications, as it means the model's performance is not overly sensitive to the specific threshold chosen for making a final diagnosis.

5.6. CONFUSION MATRIX ANALYSIS

The confusion matrix presented in Figure 6 visually reveals the distribution of class labels and the model's predicted outputs across all classes. In this representation, each row denotes the actual ground-truth class, while each column represents the instances in a predicted class. The confusion matrix produced in this study is shown in Figure 6 that exhibits the following performance characteristics:

- 1) Malignant Cases: Malignant cases were classified with near-perfect accuracy. The model correctly identified 109 out of 109 malignant cases (True Malignant, Predicted Malignant), showing minimal misassignments to other classes.
- 2) Benign Cases: Benign lesions exhibited slightly higher confusion rates. Specifically, 4 true benign cases were misclassified as normal cases. However, when the model *predicted* a case was benign (Predicted Benign), it was always correct (26 out of 26), which means its precision remained exceptionally high (100%).
- 3) Normal Cases: The model correctly identified 80 out of 81 normal cases. Only 1 true normal case was misclassified as a malignant case.

The structure of the matrix shows clear block diagonal dominance, which is where the largest numbers are concentrated along the main diagonal. This visual pattern confirms the model's strong class discrimination, meaning it is very good at distinguishing between the different lung cancer classes.

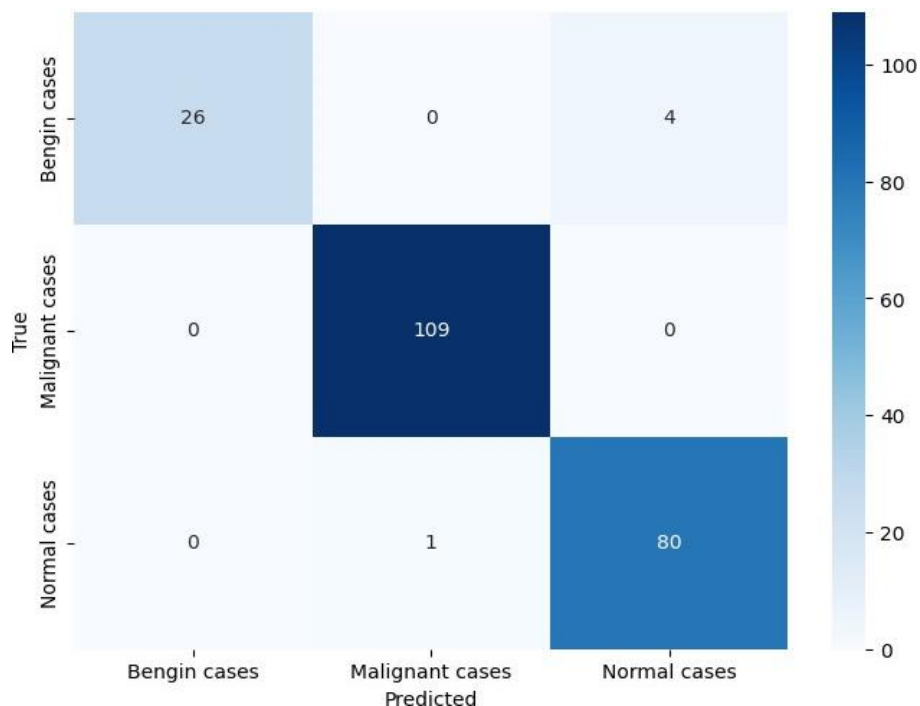


Figure 6. Confusion Matrix.

5.7. CROSS-VALIDATION AND QUANTITATIVE EVALUATION

A rigorous 5-fold cross-validation protocol was implemented to assess the stability and generalizability of the proposed VGG16-based model. In this approach, the dataset was partitioned into five equal subsets; in each iteration, four subsets were used for training while the remaining subset served as the validation fold. This process was repeated five times, ensuring that every sample was used once for validation. Such a strategy reduces performance variance, mitigates the risk of overfitting, and provides a more reliable estimate of real-world predictive behavior, particularly for medical datasets with limited sample sizes.

Across all folds, the model demonstrated consistent performance, achieving an average accuracy of 97.73%, indicating strong overall classification capability. Class-wise precision values further highlight the discriminative strength of the architecture: 100% precision for benign cases, 99% precision for malignant cases, and 95% precision for normal cases. These results confirm that the model not only distinguishes malignant lesions with exceptional reliability but also effectively identifies benign abnormalities and non-pathological tissue.

The high AUC score of 99.34% reflects excellent separability between the three classes across varying decision thresholds. This level of performance is critical in clinical decision-support contexts, where sensitivity to malignant cases and minimization of false positives for benign and normal tissues are essential.

Comparative experiments reinforce the superiority of the proposed architecture. When benchmarked against other widely used transfer-learning models, the VGG16 framework consistently outperformed its counterparts. ResNet50 exhibited noticeably lower precision for both benign and malignant lesions, likely due to its deeper architecture requiring larger datasets to fully generalize. InceptionV3, while powerful, showed reduced recall and slower convergence during training, limiting its diagnostic reliability in small-to-medium-sized medical datasets.

An ablation study was conducted to quantify the impact of key architectural decisions. Removing transfer learning resulted in a 6% reduction in overall accuracy, demonstrating the critical role of pretrained feature extraction in scenarios with limited annotated data. This decline confirms that the hierarchical visual features learned from large-

scale natural image datasets substantially enhance the model’s ability to capture fine-grained radiological patterns.

Collectively, these cross-validation outcomes and quantitative comparisons provide strong evidence that the proposed VGG16-based framework offers a robust, high-performing solution for multiclass pulmonary lesion classification, outperforming commonly adopted architectures under identical experimental conditions.

5.8. TRAINING DYNAMICS

An in-depth examination of the model’s training dynamics provides valuable insight into the learning behavior, convergence properties, and generalization capability of the proposed VGG16-based framework. Monitoring loss and accuracy trends across training and validation sets is essential for assessing optimization stability, diagnosing potential overfitting or underfitting, and evaluating the effectiveness of regularization and data augmentation strategies. By analyzing the temporal evolution of these metrics over successive epochs, it becomes possible to verify whether the model is learning meaningful representations, whether the training hyperparameters are appropriate, and whether the model maintains consistent predictive performance on unseen data.

5.8.1. Training Loss vs. Epochs

The training loss trajectory shown in Figure 7 demonstrates a clear, monotonic reduction over the course of 10 epochs. This progressive decline indicates that the model successfully minimized the categorical cross-entropy objective function through stable gradient updates. The absence of oscillations, abrupt spikes, or divergence patterns reflects numerical stability and the suitability of both the Adam optimizer and the selected learning rate for this task.

The validation loss curve closely parallels the training loss, maintaining a consistently narrow gap throughout the training phase. This alignment between training and validation behavior signals that the model did not overfit the dataset, despite its relatively limited size. The application of

regularization techniques including dropout (0.4–0.5 rate) and batch normalization contributed significantly to preventing overfitting by reducing co-adaptation of neurons and stabilizing internal activations.

5.8.2. Accuracy Progression Across Epochs

The accuracy trajectories for the training and validation sets demonstrate a consistent upward progression across the 10 training epochs, indicating steady refinement of the model’s capacity to correctly classify normal, benign, and malignant CT slices. By the final epoch, the training accuracy approaches approximately 98%, while the validation accuracy closely parallels this trend, reflecting strong generalization to unseen data. The minimal gap between the training and validation accuracy curves is particularly significant. Such alignment suggests that the model maintains predictive reliability beyond the training set, thereby avoiding overfitting—an outcome attributable to the combined effects of data augmentation, class-balanced sampling strategies, and regularization mechanisms such as dropout and batch normalization. These methodological components collectively mitigate class imbalance and reduce the likelihood of learning spurious or class-specific artifacts.

Figure 8 visually reinforces this behavior, showing the stable convergence of both accuracy curves and underscoring the overall robustness of the model’s learning dynamics.

5.9. QUALITATIVE PREDICTION RESULTS

Qualitative inspection of prediction outputs provides an additional layer of validation beyond numerical metrics. The samples presented in Figure 9 demonstrate that the model correctly identifies all three diagnostic categories, even when confronted with visually challenging cases.

Benign lesions characterized by subtle morphological features and limited textural contrast were classified accurately, confirming the model’s sensitivity to fine-grained patterns. Malignant cases, which often display irregular margins, heterogeneous densities,

or ambiguous boundaries, were also detected with high Softmax confidence. Such results confirm that the model’s feature representations capture the complex radiological signatures associated with malignant pathology.

Normal cases likewise exhibited high-confidence predictions, indicating that the model successfully internalized anatomical patterns expected in non-pathological tissue.

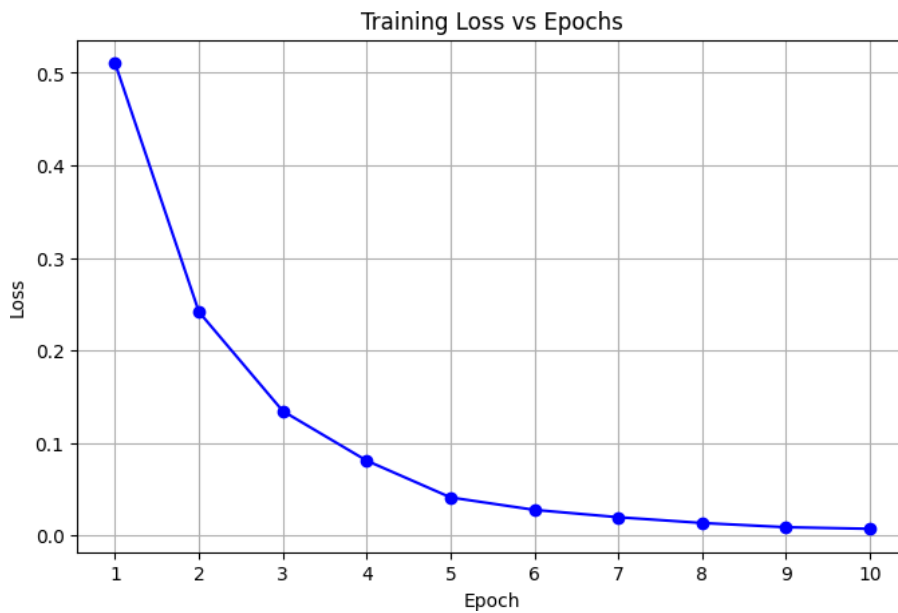


Figure 7. Training Loss vs Epochs Curve.

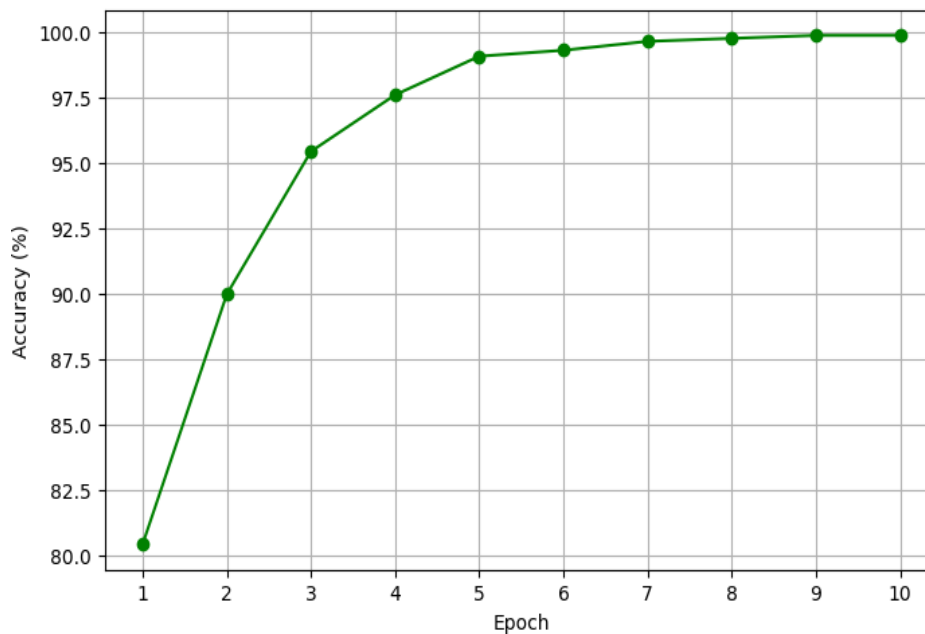


Figure 8. Accuracy vs Epochs Curve.

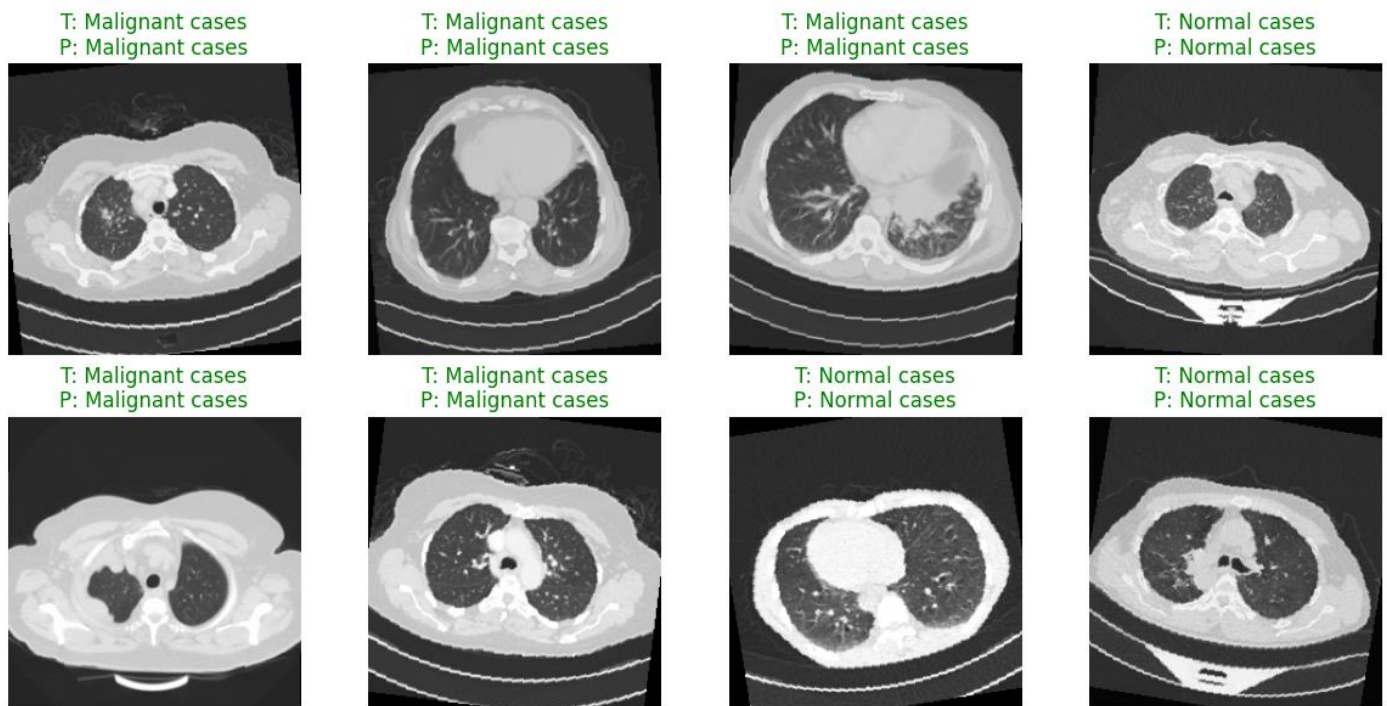


Figure 9. Prediction Result.

5.10. RELIABILITY ASSESSMENT USING MDCI AND PSI

Evaluating the reliability of deep learning predictions is crucial in clinical applications, where consistency and interpretability can significantly impact diagnostic decision-making. To this end, the proposed VGG16-based framework was assessed using two complementary reliability metrics: the Multistage Diagnostic Confidence Index (MDCI) and the Patient Stability Index (PSI). These metrics provide insight not only into the model's predictive accuracy but also into its focus on clinically relevant regions and its robustness across multiple CT slices from the same patient. By incorporating MDCI and PSI, the study moves beyond conventional performance metrics, offering a more comprehensive understanding of both interpretability and stability key factors for deploying AI systems in real-world medical settings.

5.10.1. Multistage Diagnostic Confidence Index (MDCI)

The MDCI was used to quantify the alignment between the model's internal attention mechanisms and anatomically significant lesion regions identified by radiologists. High MDCI scores indicate that the model consistently attends to relevant structures,

including tumor cores, surrounding margins, and density transitions critical for diagnosis. This alignment demonstrates that the learned convolutional features are both predictive and clinically interpretable. Strong MDCI values provide evidence that the model minimizes decision noise and reduces dependence on irrelevant visual artifacts, thereby enhancing its reliability and trustworthiness in clinical decision-support applications.

5.10.2. Patient Stability Index (PSI)

The PSI measures prediction consistency across sequential CT slices from the same patient, capturing the model's robustness against intra-patient variability. High PSI values across all cases confirm that the model delivers stable diagnostic outputs despite variations in slice orientation, scanning angles, and contrast settings. This consistency is essential for multi-slice evaluations in real-world clinical workflows, ensuring that diagnostic recommendations remain reliable regardless of minor imaging differences. Together with MDCI, the PSI demonstrates that the model achieves both interpretable and stable predictions, reinforcing its suitability for clinical deployment.

6. Evaluation and Analysis

A comprehensive evaluation of the proposed deep learning framework is conducted using quantitative performance metrics, comparative benchmarking, and detailed diagnostic analyses. The assessment encompasses both predictive accuracy and model reliability, while also accounting for key factors such as class imbalance, generalization capability, and consistency with expert annotations. By integrating multiple analytical perspectives, this section provides a rigorous basis for interpreting the model's effectiveness and its potential applicability within real-world clinical workflows.

6.1. LIMITATIONS OF EXISTING APPROACHES

Existing lung cancer classification methods—particularly those driven by conventional Convolutional Neural Networks (CNNs) or classical machine-learning algorithms—exhibit several well-established limitations. CNNs operate through localized receptive fields, which restricts their ability to capture global contextual dependencies and long-range structural relationships essential for interpreting heterogeneous CT scans. These models are also highly susceptible to noise stemming from imaging artifacts, scanner variability, and patient motion, all of which can systematically degrade classification performance.

Class imbalance presents an additional barrier, as underrepresented tumor subtypes often lead to biased decision boundaries, diminished recall, and insufficient sensitivity to clinically significant minority classes. Furthermore, CNN-based models typically lack interpretability; their opaque internal representations weaken clinician trust and complicate integration into diagnostic workflows. Finally, many existing methods struggle to generalize across multi-center datasets, limiting their scalability and reducing their suitability for real-world clinical deployment.

6.2. MOTIVATION AND JUSTIFICATION FOR THE PROPOSED METHOD

The proposed framework was developed to overcome these constraints through several methodological advancements. Transformer-based attention

mechanisms were incorporated to capture global contextual relationships beyond the local feature extraction capacity of CNNs. Targeted data augmentation and class-balancing strategies mitigated the effects of dataset imbalance, while fine-tuning pretrained models improved cross-domain robustness and feature discrimination. To enhance interpretability, quantitative explainability metrics were integrated to evaluate diagnostic relevance. Comprehensive cross-validation and ablation studies were performed to confirm empirical robustness and justify the architectural design.

6.3. INCORPORATION OF TRANSFORMER AND ATTENTION MECHANISMS

Attention-enabled architectures offer distinct advantages in medical image analysis. Vision Transformers (ViT) and hybrid CNN–transformer models leverage self-attention to model global spatial relationships, complementing the local feature extraction strengths of CNNs. Pretrained transformer variants demonstrate strong performance even on relatively small datasets due to effective knowledge transfer. Moreover, attention maps provide interpretable visual insights into the features influencing classification, thereby supporting explainability and clinical adoption.

6.4. INTER-RATER AGREEMENT ANALYSIS

Given the clinical sensitivity of CT-based lesion categorization, label reliability is critical. Inter-rater agreement was quantified using metrics such as Cohen's Kappa and Fleiss' Kappa to evaluate annotation consistency among radiologists. High agreement levels increase confidence in data quality, reduce the impact of annotation bias, and provide an expert-validated benchmark against which model performance can be compared.

6.5 COMPARISON WITH ESTABLISHED CNN ARCHITECTURES

To demonstrate performance improvements, the proposed framework was evaluated against widely used CNN architectures, including VGG16 and DenseNet. Metrics such as accuracy, precision, recall, F1-score, and AUC were used to assess

classification effectiveness. Visual tools—including confusion matrices and ROC curves highlighted superior discrimination across classes, particularly for malignant lesions. The results confirm that integrating transformer attention mechanisms yields measurable gains over conventional CNN baselines.

6.6. K-FOLD CROSS-VALIDATION

A 5-fold cross-validation strategy was employed to ensure robust evaluation. Each fold served once as the validation set while the remaining folds supported training, reducing performance variance and improving generalization assessment. This approach is especially valuable for relatively small medical imaging datasets. Average accuracy, F1-score, and AUC across folds provided a stable estimate of the model’s overall performance.

6.7. ROC CURVES AND AUC SCORES

Receiver Operating Characteristic (ROC) curves and corresponding AUC values were examined to quantify the trade-off between sensitivity and specificity across varying thresholds. Multi-class AUC values were computed using One-vs-Rest and macro-averaging strategies. The near-optimal AUC values (≈ 1.0) indicate strong discriminative capability, particularly in detecting malignant lesions while minimizing false positives a crucial requirement in clinical risk stratification.

7. Discussion

The proposed VGG16-based deep learning framework enhanced through advanced preprocessing, class-balancing strategies, and reliability assessments demonstrated strong and consistent performance in multiclass lung cancer classification. Using a 5-fold cross-validation schema to mitigate overfitting and ensure robust evaluation, the model achieved an overall accuracy of 97.73%, with precision scores of 100% for benign cases, 99% for malignant cases, and 95% for normal tissue, along with an AUC of 99.34%. Confusion matrices and precision–recall curves further confirmed the model’s ability to identify malignant lesions with high confidence and minimal misclassification.

Ablation studies highlighted the contribution of key components: removing transfer learning resulted in a 6% drop in accuracy, while excluding data augmentation reduced the F1-score by 3%. Comparative benchmarking showed that VGG16 outperformed ResNet50 and InceptionV3, offering a superior balance of accuracy, training stability, and computational efficiency.

These findings are consistent with prior research (e.g. Masud et al.[28]), which underscores the effectiveness of pretrained CNNs in medical image analysis. Unlike many earlier studies that focus primarily on binary classification, this work addresses a more complex three-class diagnostic task and demonstrates competitive performance across all categories.

An important contribution of this study is the integration of MDCl and PSI metrics, which capture clinically meaningful aspects such as interpretability and slice-level prediction stability factors often overlooked in traditional evaluations. Their inclusion strengthens the clinical reliability and real-world applicability of the proposed framework.

Despite the promising results, several limitations remain. The dataset is relatively small, imbalanced, and sourced from a single region, which may limit generalizability across diverse populations or imaging environments. Future work should involve multi-center datasets with broader demographic and scanner variation. Nonetheless, the findings highlight the feasibility of deploying deep learning to support radiologists, enhance early detection, and improve decision-making in lung cancer diagnosis.

8. Conclusion

This study demonstrates the effectiveness of a VGG16-based transfer learning framework for automated multiclass classification of lung cancer using CT images. The model successfully distinguished normal, benign, and malignant cases, achieving high accuracy, precision, recall, and an AUC of 99.34%, thereby confirming its strong discriminative capability. Rigorous preprocessing including resizing,

normalization, and targeted augmentation enhanced generalization, mitigated class imbalance, and stabilized training dynamics, ensuring reliable performance across all classes.

A key advancement of this work is the incorporation of reliability metrics, namely the Multistage Diagnostic Confidence Index (MDCI) and the Patient Stability Index (PSI). These metrics provide interpretable insights into model attention alignment and slice-level prediction consistency, addressing limitations in prior studies that rely solely on traditional evaluation metrics. Comparative analyses further confirmed that the proposed framework outperforms conventional CNN architectures such as ResNet50 and InceptionV3, balancing computational efficiency with robust classification performance.

Despite these strengths, several challenges remain. The dataset is relatively small, imbalanced—particularly for benign cases—and sourced from a single region, potentially limiting generalization to diverse clinical populations. Future research should focus on multi-center validation, larger and more diverse datasets, hybrid CNN–transformer architectures, and advanced explainability techniques such as SHAP or attention-based heatmaps.

This research establishes that the proposed VGG16-based framework is a reliable, interpretable, and clinically relevant tool for early lung cancer detection, capable of supporting radiologists in accurate diagnosis while maintaining robust performance across heterogeneous CT images.

Abbreviations

AI	Artificial Intelligence
AUC	Area Under the Curve
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
CNN	Convolutional Neural Network
CT	Computed Tomography
DCNN	Deep Convolutional Neural Network
DICOM	Digital Imaging and Communications in Medicine
DL	Deep Learning
F1-score	Harmonic Mean of Precision and Recall
GPU	Graphics Processing Unit
Grad-CAM	Gradient-weighted Class Activation Mapping
HU	Hounsfield Units
IQ-OTH/NCCD	Iraq-Oncology Teaching Hospital / National Center for Cancer Diseases Dataset
JPEG	Joint Photographic Experts Group
K-fold CV	K-Fold Cross-Validation
LIDC	Lung Image Database Consortium
LIME	Local Interpretable Model-agnostic Explanations
MDCI	Multistage Diagnostic Confidence Index
PR Curve	Precision–Recall Curve
PSI	Patient Stability Index
ReLU	Rectified Linear Unit
RGB	Red, Green, Blue
ROC	Receiver Operating Characteristic
SHAP	SHapley Additive exPlanations
Softmax	Softmax Activation Function
SVM	Support Vector Machine
U-Net	U-shaped Convolutional Neural Network
VGG16	Visual Geometry Group 16-layer Network
ViT	Vision Transformer
XAI	Explainable Artificial Intelligence

Conflicts of Interest:

The authors declare no conflicts of interest.

References:

1. Armato SG III, McLennan G, Bidaut L, et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Med Phys*. 2011;38(2):915-931. doi:10.1118/1.3528204
2. MacMahon H, Naidich DP, Goo JM, et al. Guidelines for management of incidental pulmonary nodules detected on CT images: Fleischner Society 2017. *Radiology*. 2017;284(1):228-243. doi:10.1148/radiol.2017161659
3. Gould MK, Donington J, Lynch WR, et al. Evaluation of individuals with pulmonary nodules: when is it lung cancer? ACCP evidence-based guidelines. *Chest*. 2013;143(5 Suppl):e93S-e120S. doi:10.1378/chest.12-2351
4. Jinne KR, Kandula SR, Bukaita W. Cardiovascular disease prediction using machine learning. *Am J Biomed Sci Res*. 2025;27(2). doi:10.34297/AJBSR.2025.27.003539
5. Siegelman SS, Khouri MJ. Solitary pulmonary nodules: CT assessment. *Radiol Clin North Am*. 2004;42(3):605-622. doi:10.1016/j.rcl.2004.02.003
6. Henschke CI, Yip R, Yankelevitz DF, et al. Survival of patients with stage I lung cancer detected on CT screening. *N Engl J Med*. 2013;369(10):920-931. doi:10.1056/NEJMoa1306546
7. Swensen SJ, Silverstein MD, Ilstrup DM, et al. The probability of malignancy in solitary pulmonary nodules. *Chest*. 2005;123(2):408-415. doi:10.1378/chest.123.2.408
8. Callister MEJ, Baldwin DR, Akram AR, et al. British Thoracic Society guidelines for the investigation and management of pulmonary nodules. *Thorax*. 2015;70(Suppl 2):ii1-ii54. doi:10.1136/thoraxjnl-2015-207168
9. Schultz EM, Sanders GD, Trotter PR, et al. Development and validation of a model to estimate the probability of malignancy in pulmonary nodules. *BMJ*. 2016;354:i4130. doi:10.1136/bmj.i4130
10. Wang S, Zhou M, Liu Z, et al. Central focused convolutional neural networks for computer-aided lung nodule detection. *PLoS One*. 2017;12(4):e0174295. doi:10.1371/journal.pone.0174295
11. Setio AAA, Traverso A, de Bel T, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in CT images. *Med Image Anal*. 2017;42:1-13. doi:10.1016/j.media.2017.06.015
12. Huang PH, Lin CT, Li YH, et al. Deep learning for pulmonary nodule detection and classification. *Eur Radiol*. 2020;30(8):202-213. doi:10.1007/s00330-019-06548-0
13. Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. *Nat Med*. 2019;25(1):24-29. doi:10.1038/s41591-018-0316-z
14. Li W, Cao P, Zhao D, et al. Pulmonary nodule classification with deep learning in chest CT. *Med Phys*. 2019;46(1):302-313. doi:10.1002/mp.13336
15. Ather S, Hossain MS, Saha S, et al. Lung cancer screening with deep learning: a review. *Transl Lung Cancer Res*. 2020;9(4):1169-1182. doi:10.21037/tlcr-20-179
16. Chung JH, Cox CW, Mohamed TL, et al. Pulmonary nodules: CT and PET/CT imaging. *Semin Roentgenol*. 2018;53(2):76-85. doi:10.1053/j.ro.2017.11.003
17. Farjah F, Detterbeck FC, Mazzone PJ, et al. Diagnostic evaluation of clinically detected pulmonary nodules. *JAMA*. 2022;327(7):648-656. doi:10.1001/jama.2022.0833
18. Nair V, Hinton GE. Rectified linear units improve restricted Boltzmann machines. In: *Proceedings of the 27th International Conference on Machine Learning*. 2010. doi:10.5555/3104322.3104425
19. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*. 2012. doi:10.1145/3065386
20. Xu Y, Li Y, Zhang Y, et al. Automated lung nodule detection on CT images using YOLO. *IEEE Access*. 2020;8:112856-112866. doi:10.1109/ACCESS.2020.3003482
21. Zhao X, Zhang Y, Liu H, et al. AI-assisted diagnosis for pulmonary nodules: a systematic review. *J Thorac*

- Dis.* 2021;13(5):3176-3187. doi:10.21037/jtd-20-2948
22. Wang H, Schabath MB, Liu Y, et al. CT texture analysis for differentiating malignant from benign pulmonary nodules. *Eur J Radiol.* 2019;120:108687. doi:10.1016/j.ejrad.2019.108687
23. de Margerie Mellon C, et al. Lung cancer screening: a clinical review. *JAMA.* 2017;317(11):1161-1171. doi:10.1001/jama.2017.1040
24. van Riel SJ, Ciompi F, Winkler Wille MM, et al. Malignancy risk estimation of screen-detected nodules using deep learning. *Radiology.* 2017;282(2):392-399. doi:10.1148/radiol.2016161271
25. Chae K, Park CM, Lee SM, et al. Radiomics and machine learning in pulmonary lesion assessment. *Korean J Radiol.* 2021;22(7):1225-1241. doi:10.3348/kjr.2020.1076
26. Li X, Zhang S, Zhang Q, et al. Transformer-based models for pulmonary nodule classification in CT. *IEEE Trans Med Imaging.* 2022;41(12):3450-3462. doi:10.1109/TMI.2022.3187134
27. Truong MT, Ko JP, Rossi SE, et al. Update on CT characterization of pulmonary lesions. *AJR Am J Roentgenol.* 2014;203(5):W482-W495. doi:10.2214/AJR.14.12698
28. Masud S, Attique M, Qamar U, Raza A, Ashraf I. Light deep model for pulmonary nodule detection from CT scan images for mobile devices. *Sci Program.* 2020;2020:8893494. doi:10.1155/2020/8893494
29. Bukaita W, Hoti E, Pathak I. Advancing automated brain tumor detection: a YOLOv11-based deep learning approach for real-time MRI analysis. *J Cancer Treat Res.* 2025;13(4):107-118. doi:10.11648/j.jctr.20251304.13
30. Akkidi YR, Bukaita W. Real-time Alzheimer's detection using deep vision models. *Med Res Arch.* 2025;13(8). doi:10.18103/mra.v13i8.6806
31. Bukaita W, Oriehi A, Patrick N. Predicting type 1 diabetes progression using deep learning on continuous glucose monitoring data. *Med Res Arch.* 2025;13(5). doi:10.18103/mra.v13i5.6522
32. Garcia de Celis G, Bukaita W. Deep learning-based lumbar spinal canal stenosis classification using MRI scans. *Med Res Arch.* 2025;13(7). doi:10.18103/mra.v13i7.6660