ADBA Chaos and Fractals
Cutting-Edge Scientific Solutions

e-ISSN: 3023-8951
RESEARCH ARTICLE
Vol.3 / No.1 / 2026 / pp.47-53
https://doi.org/10.69882/adba.chf.2026017

Evaluation of Deep Learning Architectures for Pulmonary CT Lesion Classification Highlighting Diagnostic Performance

Ugur Bahtiyar Guven [α,1], **Yigitcan Cakmak** [α,2] **and Ishak Pacal** [α,3]
[α]Department of Computer Engineering, Faculty of Engineering, Igdir Universty, 76000, Igdir, Türkiye.

ABSTRACT Early identification of pulmonary lesions is a critical factor in enhancing patient prognosis and survival rates. This study systematically evaluates the diagnostic performance of five deep learning architectures, ConvNeXt Base, ResNet50, EfficientNetV2 Small, InceptionV4, and Xception, for the three-class categorization of Computed Tomography (CT) scans into Benign, Malignant, and Normal categories. Utilizing the public IQ OTH NCCD dataset, we applied a transfer learning approach with ImageNet weights, complemented by a robust training pipeline incorporating dynamic data augmentation and early stopping to mitigate overfitting and ensure model generalization. Model efficacy was rigorously assessed on an independent test set using accuracy, precision, recall, and F1-score metrics. Experimental results indicate that InceptionV4 emerged as the most reliable architecture, achieving an overall accuracy of 0.988 and a macro-averaged F1-score of 0.976. Notably, this model demonstrated perfect sensitivity for the pathologically critical malignant class, achieving a recall rate of 1.00, thereby prioritizing clinical safety. These findings confirm that advanced neural networks can serve as dependable secondary opinion systems for clinicians. Given its superior sensitivity and balanced diagnostic profile, InceptionV4 represents a promising candidate for integration into automated lung cancer screening workflows to improve diagnostic precision.

KEYWORDS
Lung cancer
Deep learning
Computed tomography (CT)
InceptionV4
Transfer learning

INTRODUCTION

The prevalence of lung cancer as a serious public health concern is exemplified by its exceptionally high incidences and deaths (Baji *et al.* 2025; Ayesha *et al.* 2025), the World Health Organization (WHO) has estimated there are approximately 1.8 million deaths due to lung cancer annually, making it one of the most common causes of cancer death worldwide. The primary way of improving the prognosis and survival probability of patients diagnosed with lung cancer is through timely and accurate diagnosis. Unfortunately, patients are typically asymptomatic until the later stages of the disease when they may have the least amount of opportunities to receive an effective form of treatment (Mahmud *et al.* 2025; Rai *et al.* 2025; Siegel *et al.* 2025; Çakmak and Maman 2025). Other aspects of diagnostic, CT imaging of the lungs (the use of computed tomography) provides high-resolution cross-sectional images that may show slight nodules or other abnormalities that might have been overlooked due to the use of conventional radiological procedures (e.g., chest X-rays). Yet, the radiologist's manual interpretation of CT scans is extremely labour-intensive and time-consuming, and there is potential for variation in interpretation among different radiologists, which may result in both false positives and false negatives when distinguishing between malignant, benign, and normal lung tissues (Deepa *et al.* 2024; Venkatraman and Reddy 2024; Chowdhury *et al.* 2024).

The development of state-of-the-art artificial intelligence techniques for medical imaging has led to significant improvements in the ability of radiologists and pathologists to identify abnormal images using digital imaging systems (Attallah and Pacal 2026). With this progression in artificial intelligence comes a new set of challenges and opportunities for those working in the fields of oncology and radiology (Zafar *et al.* 2024; Dev *et al.* 2025; Çakmak and Pacal 2025; Çakmak 2025). In addition to developing algorithms that are trained to visually detect signs of cancerous tumors, researchers have developed methods to allow computer systems to analyze multiple factors related to a scanned body part (i.e., tissue density, blood flow, etc.). The ability for a computer system to learn through examples aids researchers in developing algorithms that produce better visual outputs (Jozi and Al-Suhail 2024; Lad *et al.* 2024; Lavanya *et al.* 2024; Ince *et al.* 2025).

Over the last few years, many different deep learning algorithms have been created by researchers, all aimed at identifying different types of lung cancer. The most common types of algorithms that researchers have developed include those based on the use of convolutional neural networks (CNNs), as well as hybrid systems and attention networks. Each of these types of algorithms falls into a particular category based on its design and purpose, such as optimized performance-oriented designs, architectural de-

Manuscript received: 2 November 2025,
Revised: 21 December 2025,
Accepted: 22 December 2025.

[1]ugurbahtiyarguven@gmail.com
[2]ygtcncakmak@gmail.com (**Corresponding author**).
[3]ishak.pacal@igdir.edu.tr

signs that depend on attention mechanisms for classification, and efficiency-oriented designs.

Ruprah *et al.* (2024) used Synthetic Minority Over-sampling Technique (SMOTE) combined with Gaussian blur preprocessing as part of their optimization and data handling solutions for addressing the common problem of class imbalance, demonstrating their optimized VGG-16 design had a high level of accuracy as a result of thorough data preprocessing. In a similar manner, Deepika *et al.* (2024) concentrated on hyper-parameter optimization using a two-stage methodology where initially a U-Net model is used to reconstruct the anatomical features while simultaneously reducing the noise before adding the tuned ShuffleNet classifier using Particle Swarm Optimization (PSO), which provided a high level of accuracy, 97.85

Recent studies have incorporated attention mechanisms into the feature enhancement area to direct the models to focus more on significant nodule features. For example, Muqeet *et al.* (2025) developed the Selective Kernel MobileNetV3 (SK-MNV3) incorporating the Selective Kernel Attention Mechanism in order to adapt dynamically the size of the receptive field. The SK-MNV3 model achieved 98.82% accuracy for identifying nodules of various sizes. Chaddad *et al.* (2025), in a parallel effort, designed ResNet+ by adding a Convolutional Block Attention Module (CBAM) to the base ResNet structure to maintain more spatial feature information; thus providing 99.25% accuracy when tested with the IQ-OTH/NCCD data.

In a nutshell, they wanted to compare state-of-the-art deep learning architectures regarding how they can simultaneously meet multiple tradeoffs such as high model accuracy, fast inference time, small model size, limited resources for deployment, etc. Akbari *et al.* (2024) have shown that self-attention mechanisms allow ViT models trained on histopathological images to capture global correlations better than traditional convolutional neural networks (CNNs). On the other hand, Shakya *et al.* (2025) compared compact models suitable for resource-constrained environments like edge computing and found EfficientNetV2-B0 to be the lightweight model with 98.64% accuracy.

Although the potential of custom CNNs and hybrid DL models for medical image analysis is evident in the literature, these methods typically have a significant amount of computation, and therefore, they have limitations in terms of reproducibility and standardization within clinical workflows (Munteanu *et al.* 2025). On the other hand, established state-of-the-art architectures, utilizing transfer learning, have excellent feature extraction and generalization capabilities, especially with small amounts of annotated medical data available (Moldovanu *et al.* 2024; Tătaru *et al.* 2025). With these observations in mind, this study proposes to conduct a thorough and rigorous systematic comparison of five deep learning architectures (ConvNeXt Base, ResNet-50, EfficientNetV2 Small, InceptionV4, and Xception) using the IQ-OTH/NCCD dataset for lung cancer classification across three classes. The major contributions of this research are...

- It provides a systematic evaluation of diverse architectural families (including standard CNNs, Inception-based, and lightweight models) to determine the most effective strategy for lung nodule characterization.
- Unlike studies focusing solely on overall accuracy, this research prioritizes class-wise performance metrics, specifically Recall and Precision, to identify models that minimize False Negatives in malignant cases.
- The study analyzes the trade-off between computational cost (parameters/GFLOPs) and diagnostic performance, identifying InceptionV4 as a superior candidate for reliable, automated clinical decision support systems.

MATERIALS AND METHODS

Dataset and Data Preprocessing

We used the publicly available IQ-OTH/NCCD Lung Cancer Dataset for the empirical evaluation of the proposed models. The IQ-OTH/NCCD Lung Cancer Dataset is a well-established benchmark for lung cancer classification. The dataset is available at (hamdalla alyasriy 2020). There are 1,097 images of CT scans in this dataset that belong to the three main categories of Benign, Malignant, and Normal. Figure 1 shows example images from each category and highlights the important morphological differences between cancerous, non-cancerous, and healthy lung tissue.

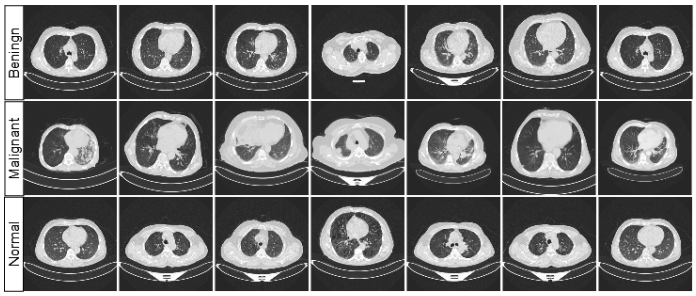

Figure 1 Sample CT Scan Images of Benign, Malignant, and Normal Lungs

In order to provide our procedures with a sound basis for our experiment, the dataset used in our studies was divided into three different datasets: training, validation and testing. The procedural split was stratified so as to ensure that the original distribution of cases within the various classes were retained across all three datasets; this allowed for the production of reliable classification models with minimal evaluation bias. As described in Table 1, our datasets used a 70/15/15 ratio to separate 767 examples for use during training; 164 examples were reserved for validation and 166 examples were held back for testing. The original distribution of classes consists of 120 examples of benign cases; 561 examples of malignant cases and 416 examples of normal cases demonstrating a large degree of class imbalance, especially regarding the absence of examples from the benign class. To compensate for the existing class imbalance and improve our models' ability to generalise and reduce the possibility of biased learning from the larger classes, advanced techniques, including data augmentation, were utilised in our methodological framework. The images were subsequently subjected to a rigorous preprocessing pipeline, including resizing and normalization, to prepare them for ingestion by the DL architectures.

Foundational Principles of Deep Learning Models

Convolutional Neural Networks (CNNs) are one of the first classes of deep learning (DL) architectures to significantly change how computer vision was done. CNNs were designed specifically to take advantage of the inherent spatial hierarchies in visual data, including images and volumes. The base of the CNN is a convolution operation wherein a set of learnable filters (also known as kernels) are slid across an input volume to perform a dot product at every spatial position between the filter weights and the corresponding input patch, producing a two-dimensional activation map that indicates the presence of specific features.

Table 1 Distribution of the Dataset into Training, Validation, and Test Sets

Classes	Original Dataset	Train (70%)	Validation (15%)	Test (15%)
Benign	120	84	18	18
Malignant	561	392	84	85
Normal	416	291	62	63
Total	1097	767	164	166

CNNs (Convolutional Neural Networks) are designed with two important rules in mind: first, that they utilize local connectivity and second, that they share parameters among all layers of the model. The purpose of these two inductive biases is to reduce the number of weights needed to train the CNN. The use of local connections allows each neuron to only connect to a small section of the previous layer's activation, or receptive field. This matches the structure of most natural images. By using the same weights across all locations in the image, the neural network is able to recognize a particular feature (e.g., a vertical line) at any place in the image. Using these two inductive biases limits the number of parameters that need to be learned in the model and helps prevent the model from overfitting.

CNN's architectural feature is its multilayer system that is structured in such a way that it learns to create a hierarchy of features. After a convolutional layer, a nonlinear activation function, like Rectified Linear Unit (ReLU), is typically utilized to help model the more complex and nonlinear relationships between the particular data inputs (features). Additionally, the convolutional layers, ReLU activation, and pooling (max pooling) layers are combined systematically throughout the CNN, this allows the CNN to produce various levels of abstraction over time, which provide a level of translational invariance (an image that has been translated has the same appearance). The initial convolutional layers learn basic primitives such as color gradients or edges, then the subsequent convolutional layers build onto these primitives to learn more complex structures and semantic information (i.e., textures, parts of objects, etc.). The final stage of this automated feature extraction process consists of a series of fully connected layers (the classification head of the CNN) which combine the high-level features and output the final feature vector, giving rise to a set of probabilities that represent the final classification outcome for the image in question.

Methodological Framework: Transfer Learning and Augmentation

In our approach to increasing our models' ability to generalize, we implemented transfer learning. We set the initial weights of our models using those learned from the prior training of the very large ImageNet dataset. Thus, our models are able to build upon the advanced feature extraction capabilities of these previous models. In order to create a model that is specific to the task of classifying pulmonary CT images, we removed the terminal classification layers of the ImageNet trained models and replaced them with a new random Dense Layer (N_Features X 3), which serves as our final Softmax classifier. This new Dense Layer has been specifically designed to fit the three-class structure of our problem. We trained our model in two distinct phases: (1) Feature Freezing - where only parameters of our new dense classifier were adjusted to fit the new data distribution and (2) Comprehensive End-to-End Fine-Tuning, where the parameters of all layers of the network were trained to adapt to very low learning rates and have been refined to classify lung nodules specifically according to their morphology.

To improve generalisation and avoid overfitting, especially considering our data set size, we've created a robust on-the-go augmentation pipeline based on the standard augmentation protocols provided via the PyTorch Image Models (timm) library. A main feature of the pipeline was to standardise the input dimensions of all images as they are re-sized to one of four (arbitrarily chosen) fixed resolutions (224x224 pixels). To preserve as many features as possible during the process of re-sizing, a random interpolation technique was used at each stage of re-sizing, where one of three different methods (bicubic, bilinear, or one of several other resampling methods) was randomly selected to use for each image. We instilled geometric invariance in our augmentation strategy through the use of Random Resized Cropping (scale between 0.08 and 1.0); as well as the use of Random Horizontal Flipping with probabilistic random selection (p=0.5). To eliminate photometric variability, which occurs in medical imaging due to differences in scanner protocol, we added Color Jittering (variance factor of 0.4), which introduced a stochastic element to the brightness, contrast, and saturation of input data. While label-mixing regularization techniques are common in natural image classification, we purposely did not employ them in this study (label-mixing coefficients set to 0.0) because we wished to maintain strict anatomic fidelity and the clear definition of the anatomic boundaries defining pulmonary nodules critical for accurate radiological evaluation (Mumuni *et al.* 2024; Wang *et al.* 2024).

Experimental Configuration and Training Procedure

In order to provide complete transparency regarding the methodology used in our research and also to allow for repeatability of the results, we conducted all of our experiments in a clearly defined, standardized computer environment. The pipeline for both training and evaluating the models was completed on a powerful workstation that contained an Intel® Core™ Ultra 7 265K CPU, 32.0 GiB of RAM, and a NVIDIA GeForce RTX™ 5090 (24GB) GPU for CUDA acceleration. The operating system on the workstation was Ubuntu 24.04.3 LTS (with Linux kernel 6.14.0-35-generic). The deep learning (DL) environment was created with Python (v3.10) using the PyTorch (v2.9.0) framework. Main libraries for data handling, training, and analysis included torchvision (v0.24.0), numpy (v2.3.4), Pillow (v12.0.0), scikit-learn (v1.7.2 - for calculating metrics), and the PyTorch Image Models (timm) library for creating and using models and training procedures.

Each individual model was trained according to the predetermined configuration parameters outlined in this document. The Stochastic Gradient Descent (SGD) Optimizer (option: sgd) was used with a momentum of 0.9 and a weight decay value of 0.000020000 (2.0e-05). A learning rate schedule using a Cosine Annealing rate scheduler (option: cosine was used to

schedule the learning rate from an initial value of 0.000010000 (warmup_learning_rate) through five successive epochs (the warmup epoch) until reaching the base learning rate. We trained every model for a maximum of 300 Epochs (total_epoch: 300), where each Epoch consists of 16 training batches (batch_size: 16). Each trained model also used label smoothing with a smoothing factor of 0.1 (smoothing: 0.1), which helps to ensure that the model is well-calibrated with respect to the expectations of its input and output data. In addition to these training parameters, we employed an early stopping strategy to prevent overfitting and discover the ideal model state. This method monitors the model's validation loss during training and terminates the training process whenever validation loss does not decrease for ten Epochs (patience_epochs: 10). When stopping, we return to the best state of the model (i.e., the best-trained version of the model) by relying on the lowest recorded validation loss.

Performance Evaluation Metrics

We performed a thorough evaluation of Model Performance on a separate Testing Set using a battery of Evaluation Metrics. Although the Aggregate Accuracy provides an overall measure of how accurate the results are, it can be deceptive in cases where there is a class imbalance. Thus, we utilized Precision, Recall and the F1-Score for a more in-depth analysis of the Models Performance. Precision is a way to express the proportion of positive predictions that are accurate; Precision is an indicator of how trustworthy a Model is by showing what portion of the positive predictions are correct. Recall (which is also referred to as Sensitivity) expresses how well a Model can detect all of the true positives in a sample. The F1-Score is defined as the harmonic mean of Precision and Recall and thus provides a composite measure of a Model Performance. The mathematical equations can be found below:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Here, TP, TN, FP, and FN denote True Positives, True Negatives, False Positives, and False Negatives, respectively. For our multiclass context, these metrics were computed independently for each class and then macro-averaged to yield a single, aggregate score that reflects the model's broadly performance across all categories.

EXPERIMENTAL RESULTS

We benchmarked five distinct architectures (ConvNeXt Base, ResNet50, EfficientNetV2 Small, InceptionV4, and Xception) on the IQ OTH NCCD dataset to verify their diagnostic utility. To guarantee an unbiased assessment, we trained each network within an identical computational environment using a fixed set of hyperparameters. Table 2 provides a numerical summary of the performance metrics, whereas Figure 2 depicts the confusion matrices to reveal category specific classification patterns.

As presented in Table 2, the experimental results reveal significant variations in performance across the tested architectures. Among the evaluated models, InceptionV4 demonstrated the most robust performance regarding clinical safety, achieving the highest overall accuracy of 0.988 and a macro-averaged F1-score of 0.976. This performance notably surpasses the baseline ResNet-50 model, which yielded the lowest accuracy of 0.939 and an F1-score of 0.870. While computational efficiency is a critical factor for clinical deployment, a clear trade-off was observed between model complexity and diagnostic precision. Although EfficientNetV2 Small demonstrated remarkable efficiency with the lowest parameter count (20.18M) and computational cost (5.419 GFLOPs), its accuracy (0.969) remained lower than that of the more complex InceptionV4. Consequently, despite its higher computational demands (12.245 GFLOPs), InceptionV4 provided the most robust balance between sensitivity and specificity, justifying its prioritization for medical scenarios where diagnostic accuracy is paramount.

Addressing the critical concern of class imbalance inherent in the dataset, where Benign cases are significantly underrepresented compared to Malignant and Normal cases, we analyzed the per-class performance using the confusion matrices illustrated in Figure 2. Typically, models trained on imbalanced data exhibit a bias toward the majority class, resulting in poor sensitivity for the minority category. However, our results provide ex-post validation of the applied on-the-fly augmentation strategy. As evidenced by the confusion matrices in Figure 2, the InceptionV4 model demonstrated remarkable calibration, correctly classifying 17 out of 18 benign cases, yielding a 0.944 recall for this minority class. This confirms that the dynamic augmentation pipeline effectively counteracted the data imbalance, allowing the model to learn distinctive features for underrepresented benign nodules without overfitting to the majority 'Malignant' or 'Normal' categories.

The superior performance of InceptionV4 as depicted in Figure 2, particularly in its ability to resolve the subtle morphological differences between benign nodules and normal tissue, can be attributed to its specialized architectural design. Unlike standard serial CNNs such as ResNet-50 or Xception that utilize fixed-size kernels, the Inception architecture employs multiple kernel sizes (1×1, 3×3, 5×5) simultaneously within each inception block. This multi-branch structure allows the network to capture multi-scale spatial features effectively, processing both the fine-grained textural details characteristic of small benign nodules and the broader morphological structures of larger malignant masses in parallel. This capability is particularly advantageous for analyzing lung CT scans where nodule size and texture variance are key diagnostic indicators, enabling InceptionV4 to outperform architectures restricted to single-scale processing in this specific application.

Visual Validation of Model Focus with Grad-CAM++

One of the most significant barriers to the clinical adoption of deep learning (DL) systems is their inherent "black-box" nature, which often obscures the underlying decision-making process. In the high-stakes environment of pulmonary oncology, achieving high statistical accuracy is insufficient; clinicians must be able to verify that the model is making its predictions based on relevant pathological features rather than spurious artifacts. To address this requirement for clinical transparency and foster trust, we utilized Grad-CAM++ to generate saliency maps for our top-performing architecture, InceptionV4. This visualization technique produces class-specific heatmaps that highlight the localized regions of the CT scan that most heavily influenced the network's final classification. By mapping these gradients back to the input image, we can qualitatively assess whether the model's focus aligns with established radiological indicators of benignity and malignancy.

As illustrated in Figure 3, the Grad-CAM++ visualizations provide compelling evidence of InceptionV4's diagnostic reliability.

Table 2 Comparison of Performance Metrics and Computational Complexity of the Models

Models	Accuracy	Precision	Recall	F1-score	Params	GFLOPs
ConvNeXt-Base	0.969	0.945	0.933	0.939	87.57M	30.707
ResNet-50	0.939	0.893	0.854	0.870	23.51M	8.263
EfficientNetV2 Small	0.969	0.957	0.920	0.936	20.18M	5.419
Inception V4	0.988	0.976	0.976	0.976	41.15M	12.245
Xception	0.975	0.980	0.925	0.948	20.81M	9.194

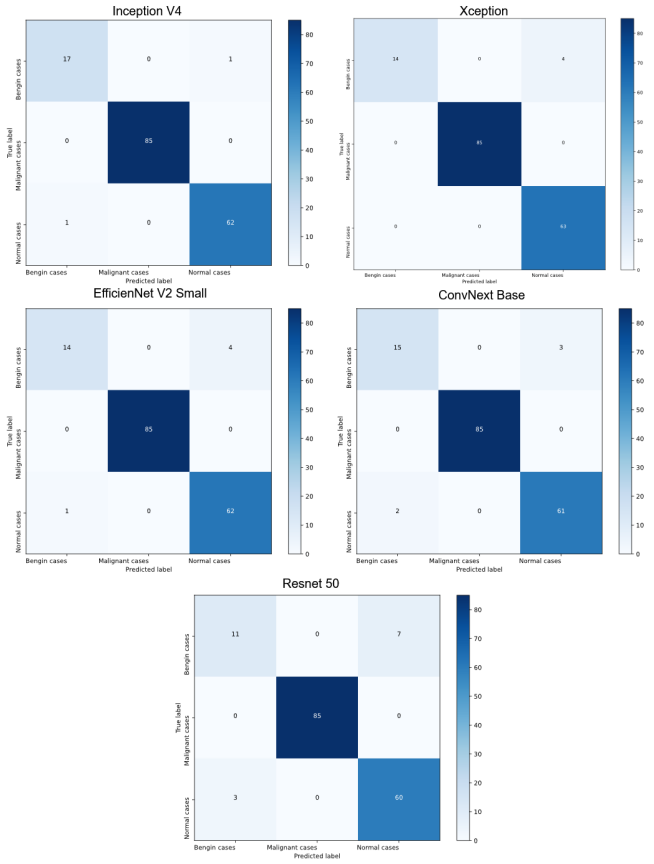

Figure 2 Confusion matrices comparing the class-wise prediction performance of the five DL models (InceptionV4, Xception, EfficientNetV2 Small, ConvNeXt Base, and ResNet-50) on the test dataset. The matrices highlight the number of True Positives (diagonal) and misclassifications (off-diagonal) for Benign, Malignant, and Normal classes.

For correctly classified cases, particularly the malignant ones where the model achieved a perfect recall rate of 1.00, the heatmaps demonstrate precise localization on the actual lesion or nodule. This localized attention is a direct result of InceptionV4's multi-scale architectural design, which utilizes parallel convolutional branches with varying kernel sizes (1×1, 3×3, and 5×5) to simultaneously capture both fine-grained textural primitives and broader morphological distortions. The saliency maps in Figure 3 confirm that the network successfully ignores non-informative anatomical structures, such as the chest wall or vascular segments, and instead prioritizes the subtle morphological cues that distinguish benign lesions from healthy tissue. This visual validation reinforces the quantitative results, confirming that InceptionV4 functions as a robust decision support tool by focusing on the same pathological regions a radiologist would examine.

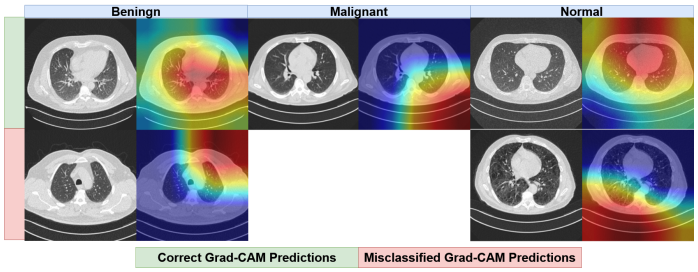

Figure 3 Visual validation of diagnostic focus using Grad-CAM++ saliency maps for the InceptionV4 architecture across Benign, Malignant, and Normal lung CT categories, illustrating both correct and misclassified instances.

DISCUSSION

The accurate characterization of pulmonary nodules represents a pivotal challenge in modern oncology where the distinction between benign lesions and malignant tumors directly dictates patient survival. While DL has emerged as a transformative force in this domain, the clinical adoption of these systems hinges not on marginal improvements in accuracy but on their reliability in preventing missed diagnoses. Within this context, our study reveals a critical insight: while overall accuracy is a standard benchmark in machine learning, the true utility of an AI system in medical imaging is defined by its safety profile. Among the diverse architectures analyzed, InceptionV4 emerged not merely as a statistical leader but as the most clinically robust candidate. It achieved an overall accuracy of 0.988 and a macro-average F1-score of 0.976. However, the defining characteristic of this performance, and the primary contribution of this work, is the model's perfect sensitivity as evidenced by a 1.00 recall rate for malignant cases. In the high-stakes domain of lung cancer screening where a missed diagnosis can drastically alter patient prognosis, the elimination of false negatives is paramount. This capability positions the proposed framework as a reliable safety net for radiologists since it prioritizes patient survival over marginal gains in general classification metrics.

This superior sensitivity is not accidental but stems from the specific inductive biases inherent in the InceptionV4 architecture.

Unlike serial backbones such as ResNet-50 which rely on fixed receptive fields, InceptionV4 utilizes parallel convolutional branches with varying kernel sizes. This multi-scale feature extraction allows the network to simultaneously resolve the fine-grained textures of small benign nodules and the gross morphological distortions typical of malignant masses. This architectural advantage effectively addresses the challenge of distinguishing subtle benign lesions from normal tissue, a task where single-scale models often falter.

We acknowledge that recent benchmarks on the IQ-OTH/NCCD dataset have reported marginally higher numerical accuracies, such as those achieved by Chaddad et al. using attention-enhanced ResNet+, or Sheikh Akbari et al. via ViT. However, the slight performance gap in our results reflects a deliberate methodological trade-off. Unlike recent studies that focus solely on maximizing accuracy through complex hybrid models, our work prioritizes architectural explainability and clinical safety. We demonstrate that the multi-scale processing of InceptionV4 effectively eliminates false negatives without the need for aggressive synthetic data generation. Consequently, we consciously excluded label-altering augmentation techniques like Mixup or Cutmix to preserve the strict anatomical fidelity of the CT images. While such techniques might artificially boost validation metrics, they risk introducing artifacts that are not clinically representative. Therefore, the perfect malignant recall achieved by InceptionV4 without synthetic pixel interpolation confirms its practical utility as a robust decision support tool.

CONCLUSION

The findings of this study highlight the necessity of rethinking conventional model evaluation practices by moving beyond isolated accuracy metrics toward a more comprehensive assessment of diagnostic safety and reliability. The demonstrated robustness and reproducibility of the proposed transfer learning framework underscore its practicality and suitability for integration into standard medical imaging workflows. By complementing radiological expertise and mitigating inter-observer variability, such automated systems have the potential to function as effective clinical decision support tools. Looking ahead, future research will focus on validating the proposed approach across multi-center datasets and incorporating explainable artificial intelligence techniques to enhance clinical trust, transparency, and interpretability.

Despite the promising results, we must acknowledge some limitations to ensure a balanced scientific interpretation of this study. A primary limitation is that the results are from a single data source, the IQ-OTH/NCCD dataset. While this dataset serves as a standard benchmark, evaluating models on a single cohort is not fully objective, as differences in CT scanner protocols, reconstruction kernels, and patient demographics across institutions can significantly impact model performance. Furthermore, while a test set size of 166 images is sufficient for statistical indices, it is not an effective test data set. This limited sample size creates the risk of statistical variance, and the observed performance differences should be tested with larger, multicenter, and diverse datasets to verify their robustness across various clinical settings. Unfortunately, due to the lack of publicly available datasets, we must accept these limitations.

Additionally, the dataset exhibits a critical class imbalance, with the "Benign" class being significantly underrepresented compared to "Malignant" and "Normal" cases. Although we employed a robust data augmentation pipeline to mitigate this issue, the confusion matrices reveal that the majority of misclassifications occurred within the benign class. While InceptionV4 handled this imbalance effectively, other models like ResNet-50 struggled to distinguish benign nodules from normal tissue, highlighting that algorithmic bias toward the majority class remains a challenge. Furthermore, this study's scope was intentionally limited to the CT imaging modality. We acknowledge that a multimodal approach, integrating data from other sources such as Positron Emission Tomography (PET) or histopathological slides, was not explored but presents a valuable direction for future research. Another limitation is the lack of explainability; as DL models function as "black boxes," high accuracy does not guarantee that the model is focusing on clinically relevant pathological features. Future research will aim to address these shortcomings by validating the top-performing InceptionV4 model on external, large-scale datasets such as LIDC-IDRI to ensure cross-center generalizability. We also plan to integrate XAI modules, such as Grad-CAM, to visualize the decision-making process and ensure alignment with radiological expertise, alongside exploring ensemble methods that combine the high sensitivity of InceptionV4 with the efficiency of lightweight models.

Ethical standard

Ethical approval was not required for this study as it relied exclusively on the publicly available 'IQ-OTH/NCCD Lung Cancer Dataset'. Since the data consists of pre-existing, anonymized CT images and does not involve direct interaction with human subjects or animals, institutional review board (IRB) certification is not applicable.

Availability of data and materials

The "IQ-OTH/NCCD Lung Cancer Dataset" used for this study is publicly available and accessible, as cited in reference (32).

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

LITERATURE CITED

Akbari, A. S., A. Kumar, B. R. Reddy, K. K. Singh, and M. Takei, 2024 Vision transformer based automated model for enhancing lung cancer classification. In *2024 IEEE International Conference on Imaging Systems and Techniques (IST)*, pp. 1–6, IEEE.

Attallah, O. and I. Pacal, 2026 Impact of magnification on deep learning approaches through comprehensive comparative study of histopathological breast cancer classification. Biomedical Signal Processing and Control **113**: 108973.

Ayesha, N., I. H. Hassan, A. R. Mirdad, and A. R. Khan, 2025 Efficientnet deep learning model for lung cancer early diagnosis from computed tomography scan images with transfer learning. Journal of Advances in Information Technology **16**.

Baji, S. R., S. B. Bagal, S. V. Chaudhari, and B. S. Agarkar, 2025 Effective diagnosis of lung cancer using pyramid quantum convolutional neural network with migrating walrus algorithm on ct scan images. Biomedical Materials & Devices pp. 1–26.

Çakmak, Y., 2025 Machine learning approaches for enhanced diagnosis of hematological disorders. Computational Systems and Artificial Intelligence **1**: 8–14.

Çakmak, Y. and A. Maman, 2025 Deep learning for early diagnosis of lung cancer. Computational Systems and Artificial Intelligence **1**: 20–25.

Çakmak, Y. and N. Pacal, 2025 Deep learning for automated breast cancer detection in ultrasound: A comparative study of four cnn architectures. Artificial Intelligence in Applied Sciences **1**: 13–19.

Çakmak, Y. and J. Zeynalov, 2025 A comparative analysis of convolutional neural network architectures for breast cancer classification from mammograms. Artificial Intelligence in Applied Sciences **1**: 28–34.

Chaddad, A., J. Peng, and Y. Wu, 2025 Classification based deep learning models for lung cancer and disease using medical images. IEEE Transactions on Radiation and Plasma Medical Sciences .

Chowdhury, A., T. Moni, A. A. N. Tushar, M. P. Hossain, and M. A. Rahaman, 2024 An improved deep learning model for early stage lung cancer detection from ct scan images. In *2024 6th International Conference on Sustainable Technologies for Industry 5.0 (STI)*, pp. 1–6, IEEE.

Deepa, V. *et al.*, 2024 Fossil_net lung cancer prediction and classification from ct images using convolution neural networks. In *2024 2nd International Conference on Computing and Data Analytics (ICCDA)*, pp. 1–5, IEEE.

Deepika, R., P. Shanmugam, K. Moorthi, P. M. Kumar, S. Swarna, *et al.*, 2024 Optimized transfer learning model for lung cancer stage classification using computed tomography images. In *2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS)*, pp. 912–917, IEEE.

Dev, S., P. S. Roy, N. Chakraborty, and R. Sarkar, 2025 Lung cancer identification from ct scans using a soft-attention enabled deep transfer learning model. In *2025 3rd International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC)*, pp. 254–259, IEEE.

hamdalla alyasriy, 2020 The iq-othnccd lung cancer dataset **1**.

Ince, S., I. Kunduracioglu, A. Algarni, B. Bayram, and I. Pacal, 2025 Deep learning for cerebral vascular occlusion segmentation: a novel convnextv2 and grn-integrated u-net framework for diffusion-weighted imaging. Neuroscience **574**: 42–53.

Jozi, N. S. and G. A. Al-Suhail, 2024 Lung cancer detection: The role of transfer learning in medical imaging. In *2024 International Conference on Future Telecommunications and Artificial Intelligence (IC-FTAI)*, pp. 1–6, IEEE.

Lad, S., B. Chafekar, and P. Bide, 2024 Lung cancer classification using deep learning: A comprehensive approach with modified convolutional neural networks. In *2024 International Conference on Computational Intelligence and Network Systems (CINS)*, pp. 1–6, IEEE.

Lavanya, G., M. Muthulakshmi, M. Latha, A. Keerthinathan, P. V. Krishh, *et al.*, 2024 Deep learning for enhanced detection and characterization of pulmonary nodules. In *2023 4th International Conference on Intelligent Technologies (CONIT)*, pp. 1–7, IEEE.

Mahmud, M. R., H. Fardin, M. I. H. Siddiqui, A. H. Sakib, and A. Al Sakib, 2025 Hybrid deep learning for interpretable lung cancer recognition across computed tomography and histopathological imaging modalities. International Journal of Science and Research Archive [Internet] pp. 1798–810.

Moldovanu, S., G. Tăbăcaru, and M. Barbu, 2024 Convolutional neural network–machine learning model: hybrid model for meningioma tumour and healthy brain classification. Journal of Imaging **10**: 235.

Mumuni, A., F. Mumuni, and N. K. Gerrar, 2024 A survey of synthetic data augmentation methods in machine vision. Machine Intelligence Research 2024 21:5 **21**: 831–869.

Munteanu, D., S. Moldovanu, G. Tăbăcaru, and M. Barbu, 2025 Influence of symmetric and asymmetric cae-cnn on colon cancer histopathological images classification. In *2025 33rd Mediterranean Conference on Control and Automation (MED)*, pp. 203–208, IEEE.

Muqeet, M. S. *et al.*, 2025 Enhancing early detection of chronic obstructive pulmonary disease using high-resolution mri and advanced deep learning techniques. In *2025 Global Conference in Emerging Technology (GINOTECH)*, pp. 1–8, IEEE.

Pacal, I. and O. Attallah, 2025 Inceptionnext-transformer: A novel multi-scale deep feature learning architecture for multimodal breast cancer diagnosis. Biomedical Signal Processing and Control **110**: 108116.

Rai, N., S. Khatri, and D. Risal, 2025 Explainable ai technique in lung cancer detection using convolutional neural networks. arXiv preprint arXiv:2508.10196 .

Ruprah, T. S., B. Regmi, S. B. Jadhav, and S. Singh, 2024 Early stage lung cancer detection using deep learning. In *2024 MIT Art, Design and Technology School of Computing International Conference (MITADTSoCiCon)*, pp. 1–6, IEEE.

Shakya, S. R., E. Ceh-Varela, and I. Sanjaya, 2025 Lung cancer classification using deep learning models for edge computing. a comparative analysis. In *2025 IEEE International Conference on AI and Data Analytics (ICAD)*, pp. 1–6, IEEE.

Siegel, R. L., T. B. Kratzer, A. N. Giaquinto, H. Sung, and A. Jemal, 2025 Cancer statistics, 2025. Ca **75**: 10.

Tătaru, I., S. Moldovanu, O.-M. Dragostin, C. L. Chiţescu, A.-S. Zamfir, *et al.*, 2025 Auto machine learning and convolutional neural network in diabetes mellitus research—the role of histopathological images in designing and exploring experimental models. Biomedicines **13**: 1494.

Venkatraman, K. and S. N. P. S. Reddy, 2024 Augmenting clinical decisions with deep learning lung cancer image abnormality segmentation. In *2024 14th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pp. 674–678, IEEE.

Wang, Z., P. Wang, K. Liu, P. Wang, Y. Fu, *et al.*, 2024 A comprehensive survey on data augmentation .

Zafar, S., J. Ahmad, Z. Mubeen, and G. Mumtaz, 2024 Enhanced lung cancer detection and classification with mrmr-based hybrid deep learning model. Journal of Computing & Biomedical Informatics **7**.

How to cite this article: Guven, U. B., Cakmak, Y., and Pacal, I. Evaluation of Deep Learning Architectures for Pulmonary CT Lesion Classification Highlighting Diagnostic Performance. *Chaos and Fractals*, 3(1), 47-53, 2026.