

Article

Lung Cancer Classification from CT Images Using Ensemble Transfer Learning of VGG16, ResNet50V2 and MobileNetV2

İshak Metehan Sis ^{1,*} , Mete Yağanoğlu ¹  and Adem Maman ² ¹ Department of Computer Engineering, Faculty of Engineering, Ataturk University, 25240 Erzurum, Türkiye; yaganoglu@atauni.edu.tr² Department of Nuclear Medicine, Ataturk University Research Hospital, Ataturk University, 25240 Erzurum, Türkiye; adem.maman@atauni.edu.tr

* Correspondence: ishak.metehansis@atauni.edu.tr

Abstract

Lung cancer remains one of the most common cancers worldwide, and how early it is caught shapes what treatment is possible and how long a patient lives. Computed tomography (CT) sits at the center of this process, and deep learning has become a standard tool for reading CT images, with strong results across medical image classification tasks. This paper proposes an ensemble approach that brings pre-trained VGG16, ResNet50V2, and MobileNetV2 together in a single framework, classifying lung CT images as benign, malignant, or normal. The dataset consists of 1946 CT images from 473 patients, collected at Atatürk University Hospital and labeled by a nuclear medicine physician. All three backbones were fully fine-tuned end-to-end, an approach that an ablation study confirmed outperformed both fully frozen and partially frozen configurations. Splits were made at the patient level throughout, so no patient appears in both training and validation. Five-fold cross-validation gave $89.42 \pm 3.21\%$ accuracy and a macro F1 of 0.865 ± 0.036 ; the same configuration, retrained on a separate patient-level split, reached 90.62% accuracy with a macro F1 of 0.874. The results indicate that ensemble transfer learning can serve as a supporting tool in lung cancer screening.

Keywords: lung cancer detection; deep learning; transfer learning; ensemble learning; image classification; VGG16; ResNet50V2; MobileNetV2

1. Introduction

In living organisms, cells multiply according to a specific order to ensure growth and tissue renewal. Cancer can be defined as a disease that arises as a result of the disruption of this order, causing cells to multiply uncontrollably and abnormally. Cancer can be caused by unhealthy eating habits, alcohol consumption, exposure to chemicals, obesity, insufficient physical activity, and various environmental factors [1]. Almost a quarter of cancer deaths worldwide are due to lung cancer.

Various medical imaging methods such as chest radiography, computed tomography (CT), and positron emission tomography (PET) are used in the diagnosis of lung cancer. Due to the shortcomings of these methods in interpreting images and automatically classifying lesions, there has been a shift towards artificial intelligence and deep learning-based approaches in this field in recent years. Accurate differentiation of findings into benign and malignant and their distinction from normal tissue is extremely important for timely and effective clinical management [2,3].



Academic Editor: Alexander Zhbanov

Received: 20 June 2026

Revised: 23 July 2026

Accepted: 28 July 2026

Published: 30 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Artificial intelligence is a technology that can mimic human intelligence in the processes of analyzing, interpreting, and making sense of large and multidimensional data sets using various algorithms and software techniques. Today, artificial intelligence applications are effectively used in many healthcare fields such as diagnosis, treatment planning, disease monitoring, and medical image analysis. In recent years, developments in the field of artificial intelligence have led to significant progress in lung cancer detection. In particular, deep learning-based algorithms have achieved high accuracy rates in the characterization of lung nodules, producing results close to or superior to the performance of expert radiologists in differentiating between benign and malignant nodules [4]. It has also been widely adapted to clinical imaging applications [5].

Deep learning is one of the most widely used machine learning algorithms. Studies using deep learning in the medical field have become quite widespread recently with the increase in data. In the literature, many studies have been conducted using deep learning algorithms for the diagnosis of lung cancer and successful results have been obtained. In addition, transfer learning, which means fine-tuning a pre-trained model on a large-scale dataset on a task-specific dataset, has emerged as a particularly practical strategy in medical imaging scenarios where labeled data is limited and data collection costs are high [6]. The scope and contributions of this study can be summarized as follows:

- The study presents an ensemble transfer learning model that combines VGG16, ResNet50V2 and MobileNetV2 to classify lung CT images into benign, malignant and normal categories.
- The dataset comprises 1946 CT images from 473 patients, directly labeled by a co-author nuclear medicine physician based on clinical evaluation. This expert-led labeling minimizes label noise, ensuring that the class boundaries learned by the model accurately reflect real clinical distinctions.
- VGG16 provides deep and sequential convolution representations; ResNet50V2 uses pre-activated residual jump connections to capture high-level features without gradient distortion; MobileNetV2 offers computationally efficient, depth-separable convolution features.
- Backbone selection was guided by architectural dissimilarity rather than intuition or accuracy. VGG19 and ResNet50, close relatives of the VGG16 and ResNet50V2 already in the ensemble, would likely add redundancy. MobileNetV2 was selected instead, offering a different representational approach.
- The extracted feature vectors are combined and passed through fully connected layers to produce class-level probability outputs.
- Data preprocessing steps include resizing the images to 224×224 pixels, preserving the RGB channels and pixel normalization.
- Data augmentation-based training diversity is used to improve generalization against unseen imaging conditions.
- Experimental results demonstrate that the ensemble model outperforms both the single-layer model and similar ensemble models.
- Rather than assuming a freezing strategy, we tested one: a systematic ablation comparing fully frozen, partially frozen, and fully open configurations under the same training conditions. Fully fine-tuning all three backbones came out on top, so that is what we used for the final model.
- A dedicated ablation study isolating the individual contribution of each backbone is included. This provides empirical support for the three-model ensemble choice rather than treating architectural complementarity as a given assumption.

The remainder of the study is structured as follows: Section 2 reviews similar studies. Section 3 describes the methodology, including dataset characteristics, preprocessing

pipeline, model architecture, and training configuration. Section 4 presents the quantitative evaluation results and comparative analysis. The final sections discuss the implications of the findings, examine the differences and advantages compared to existing studies, and summarize future research directions.

2. Related Work

The first systems aimed at automatically diagnosing lung cancer attempted to process visual criteria manually determined by medical professionals using traditional algorithms such as Support Vector Machines (SVM), Decision Trees, and k-Nearest Neighbors (k-NN); however, these methods required significant expertise and could not easily adapt to complex datasets from different sources [7,8]. The introduction of Convolutional Neural Networks (CNNs) revolutionized this field because these networks learned key details in images directly from pixels, eliminating the laborious feature extraction process done manually [9]. Subsequently, transfer learning, which uses the knowledge of pre-trained models in massive datasets, became a popular standard, especially in scenarios where finding medical images is difficult, and improved diagnostic success despite limited data [10]. Today, ensemble architectures that combine the predictions of multiple AI models are used; this minimizes the margin of error or bias of a single model, ensuring that systems provide much more consistent and reliable results [11].

Çiftçi conducted a study to determine whether a patient has cancer and, if so, whether it is benign or malignant, using a dataset of 6053 lung tomography scans. He used two different stages to divide the data into different ratios. In the first stage, VGG-16, Inception v4, and MobileNetV3 were used, and in the second stage, the AlexNet method was used. As a result of the study, it was seen that AlexNet gave a result with an accuracy of 96%, MobileNetV3 with 81%, VGG-16 with 84%, and Inception v4 with 86% [12].

Üzülmez and Çiftçi conducted a study combining deep learning methods with Uncertainty Quantization (UQ) to detect lung cancer and classify cancerous nodules as benign or malignant. The study used a dataset of 663,549 lung CT scans and applied two different methods. In the first method, the performance of different deep learning models was evaluated without using the Uncertainty Quantization approach. According to the results, the ResNet50 model achieved an accuracy rate of 96.5%, the AlexNet model 83.7%, the InceptionV3 model 84.9%, and the VGG16 model 82.6%. Furthermore, the accuracy rate of the developed 4-layer Convolutional Neural Network (CNN) model was calculated as 97.1%. In the second method, the Bayesian Neural Network-based Uncertainty Quantization method was applied to increase the predictive reliability of the models. As a result of this approach, ResNet50, AlexNet, InceptionV3 and VGG16 models achieved accuracy rates of 89%, 86%, 91% and 77%, respectively [13].

Şeker et al. conducted a study on the automatic classification of lung cancer from histopathological images. In the study, the performance of Convolutional Neural Network (CNN), Vision Transformer (ViT), and EfficientNet-B1 deep learning architectures was compared, and the LIME (Local Interpretable Model-agnostic Explanations) method was used to improve the interpretability of the model outputs. As a result of the evaluations, it was determined that ViT and EfficientNet-B1 models showed higher performance than the CNN model in F1-score, sensitivity, and precision metrics. While the most successful classification results were obtained with the ViT architecture, the most comprehensive explanations in terms of visualization and interpretation of model decisions were provided by the EfficientNet-B1 model. The findings show that transformer-based and modern deep learning architectures can be effectively used in the diagnosis of lung cancer from histopathological images [14].

Shah et al. developed a two-dimensional CNN-based deep learning ensemble to detect lung nodules. Achieving a 95% accuracy rate, they demonstrated that combining multiple models is more effective than using single networks [15].

Mikhael et al. introduced a special deep learning system they named Sybil. This model can predict future lung cancer risk with high precision from a single low-dose CT scan and demonstrates that it is possible to transition from single-timepoint imaging to longitudinal risk stratification [16].

Wankhade and Vigneshwari, focusing on the diagnosis of early-stage lung cancer, combined three-dimensional CNNs with hybrid deep learning modules. Their developed CCDC-CNN model provided a stronger distinction between benign and malignant nodules in volumetric data obtained from CT [17].

Said et al. developed a segmentation-oriented classification pipeline by combining UNETR architectures with self-supervised learning networks. This approach provided high diagnostic accuracy in lung images obtained from CT [18].

Wani et al. presented a hybrid framework they called DeepXplainer. Combining CNN-based feature extraction with XGBoost classification, this system increased clinical interpretability while maintaining predictive performance thanks to gradient-based explainability mechanisms. They achieved 97.43% accuracy with their proposed method [19].

Chae et al. applied deep learning-based texture analysis to detect interstitial lung abnormalities in large-scale CT scan datasets and demonstrated the clinical significance of texture-sensitive convolutional representations [20].

Guan et al. developed an automated PET image analysis framework utilizing differential activation filters and CNNs. This system demonstrated cutting-edge performance in pulmonary scan benchmarks [21].

Mohamed et al. proposed the hybrid algorithm EOSA-CNN, which combines an Ebola optimization-based search strategy with CNN classification. This method achieved high accuracy in CT-based lung cancer datasets [22].

Rajasekar et al. achieved significant success in the early detection of malignant tissue microstructures, particularly by using multiple deep learning models in the analysis of histopathological cross-sectional images [23].

Kusuma achieved 97.3% accuracy in early-stage lung cancer classification by proposing a hybrid CNN-RNN framework enhanced with the Pelican Optimization Algorithm [24].

Reshma et al. compared deep CNN architectures with traditional CNNs in the classification of low-dose CT images and reported 95% accuracy, highlighting the advantages of deeper network configurations [25].

Mohana Krishna and Puviarasi compared ResNet50 with InceptionV3 for CT-based lung cancer prediction, achieving 93.09% accuracy and observing that residual connections provided more stable gradient flow during training [26].

Safta and Shaffie improved pulmonary nodule diagnosis by integrating 3D-LOP local descriptive features with three-dimensional CNN models and achieved 97.84% accuracy [27].

Princy Magdaline and Ganesh Babu combined attention-gated residual U-Net segmentation with k-nearest neighbor classification and achieved 97% detection accuracy in CT image datasets [28].

Kumaran et al. proposed an explainable ensemble framework integrating VGG16, ResNet50, and InceptionV3 backbones with Grad-CAM visualization. This study, which achieved 98.18% accuracy in the IQ-OTH/NCCD dataset, demonstrated the advantages of multi-architecture feature fusion in terms of diagnostic interpretability [29].

A recent systematic review by Yang et al. published in the journal Applied Sciences (MDPI) presented a comprehensive taxonomy of deep learning-based segmentation and

classification methods in lung cancer CT images. This analysis, encompassing CNN-based, U-Net-based, and Transformer-supported architectures, confirmed that ensemble learning strategies consistently achieve top-level classification accuracy and supported the methodological approach followed in this study [30].

The key difference that sets this research apart from similar studies in the literature can be summarized as follows:

- An ensemble model consisting of VGG16, ResNet50V2, and MobileNetV2, which has not been used before in lung tumor classification, is proposed.
- Many similar studies have trained their models on the same publicly available datasets.
- The new dataset, created from 473 patients and 1946 CT images and labeled by the nuclear medicine physician involved in the study, offers a new alternative to publicly available datasets.
- Ensemble models are more competitive in terms of performance compared to single models. Moreover, they have earned a highly competitive place in the literature with their demonstrated performance.

3. Materials and Methods

3.1. Description of the Data Set

The experimental dataset used in this study consists of 1946 CT images obtained from a total of 473 patients. All images were collected under clinical conditions and, based on radiological evaluation, were classified into three diagnostic categories: benign (Benign CT), malignant (Malignant CT), and normal (Clear CT). All patients whose images were classified as malignant were diagnosed with lung cancer. Malignant cases constitute the largest share of the dataset, accounting for 41.7%, while benign and normal cases are represented at 18.3% and 40.0%, respectively. Table 1 shows the distribution of CT images in the dataset.

Table 1. Class-wise distribution of CT images in the dataset.

Class	Patients	Images	Avg. Images /Patient	Image (%)
Benign	131	356	2.72	18.3%
Malignant	245	812	3.31	41.7%
Normal	97	778	8.02	40.0%
Total	473	1946	4.11	100%

Labelling followed clinical criteria applied consistently across the dataset by the nuclear medicine physician co-author. A nodule was classified as benign when it showed no Fluorodeoxyglucose (FDG) uptake on PET/CT and no change in size across consecutive scans six months apart, with a minimum detectable size of 3 mm. A nodule was classified as malignant either through histopathological confirmation of primary lung cancer or pulmonary metastasis, or when consecutive scans three months apart showed both FDG uptake and growth in size.

All CT images used in this study were fully anonymized at the point of archival extraction, in line with the ethics approval and consent waiver for this retrospective study. As a result, individual patient-level information such as age and sex was not available and could not be recovered. Scanner-level acquisition parameters, however, were retained: all images were acquired using a GE Healthcare Discovery IQ PET/CT scanner (GE Medical Systems, Waukesha, WI, USA), with a tube voltage of 120 kVp, tube current of 227 mAs, gantry rotation time of 0.70 s, and a reconstructed slice thickness of 2.5 mm. The dataset was

also collected from a single institution (Ataturk University Hospital), which may introduce center-specific characteristics related to scanner type and imaging protocol.

Every image in the dataset carries a unique, de-identified patient code, allowing all images from the same patient to be grouped together regardless of their diagnostic category. This patient-level grouping ensures that images from a single patient are never split across different subsets in any experiment, preventing data leakage as detailed in Section 3.2.

3.2. Data Preprocessing

In order to ensure input consistency and compatibility with the selected architectures, a standard preprocessing was applied to all images prior to model training. The images were loaded from the relevant class directories and converted to three-channel RGB format, regardless of the original color space. Then, each image was resized to a resolution of 224×224 pixels, which is compatible with the input requirements of the VGG16, ResNet50V2 and MobileNetV2 architectures. This resolution is also considered sufficient for preserving diagnostically critical structural details.

$$I_{\text{resize}} = \text{resize}(I_{\text{original}}, 224 \times 224)$$

After spatial normalization, the pixel density values were scaled to the intervals $[0, 1]$ by dividing by 255, which is the maximum value that can be represented in 8-bit images. This step provides a consistent feature scaling between the inputs and contributes to the stable stabilization of gradient dynamics during the training process.

$$I_{\text{normalize}} = I_{\text{resize}}/255$$

All data partitioning in this study was performed at the patient level rather than the image level. Because a single patient can contribute multiple CT slices, an image-level split allows near-identical images from the same patient to appear in both the training and validation sets, which artificially inflates measured performance. To avoid this, patient identifiers were used as the grouping variable for every split reported in this study. The two exceptions are the replication of Kumaran et al. [29] and the IQ-OTH/NCCD control experiment reported in Section 4; both use image-level splits for the reasons given there.

Model configuration was selected using five-fold GroupKFold cross-validation over the full dataset, with folds constructed so that all images belonging to a given patient fall entirely within a single fold. This procedure was used to evaluate the stability of candidate configurations across different patient partitions, since single-split performance rankings were found to vary substantially between splits. The selected configuration was then retrained on a separate, patient-level 80/20 split, and the results reported in Section 4 correspond to the validation portion of that split. In both the cross-validation procedure and the final split, no patient contributes images to more than one subset.

Class labels were converted to one-hot encoding to be compatible with the three-class softmax output layer and the categorical cross-entropy loss function.

3.3. Data Augmentation

To reduce the risk of overfitting and improve the generalizability of the model to unfamiliar imaging conditions, data augmentation transformations were applied to training samples at each epoch. Augmentation was limited to the training subset; validation samples were not subjected to any geometric transformations to ensure an unbiased performance evaluation. The implemented augmentation procedures are as follows:

- Random rotation in the range of $\pm 15^\circ$ to simulate patient positioning differences during CT imaging (rotation_range = 15).

- Horizontal shift of up to 10% of image width to account for lateral variations in lung field positioning (`width_shift_range = 0.1`).
- Vertical shift of up to 10% of image height to reflect superior-inferior anatomical variability (`height_shift_range = 0.1`).
- Horizontal flip to provide bidirectional symmetry variation (`horizontal_flip = True`).
- Zoom change in the range of $\pm 10\%$ to simulate image magnification and field of view variations (`zoom_range = 0.1`).
- Nearest neighbor fill for boundary regions resulting from geometric transformations (`fill_mode = 'nearest'`).

The batch size was fixed at 32 for both generators, and only the training batches were remixed at each epoch.

To maintain data integrity, data augmentation was restricted strictly to the training set following the split. Consequently, the validation set contained only original, unaltered images to ensure an unbiased evaluation of the model's generalization capability.

3.4. Model Architecture

The proposed classification model is based on an ensemble architecture that combines three pre-trained models (VGG16, ResNet50V2, and MobileNetV2) working on an input image. Each model functions as an independent feature extractor capturing complementary features of lung tissue morphology at different levels of abstraction.

Developed by the Oxford University Visual Geometry Group, VGG16 sequentially applies 3×3 convolution filters that gradually deepen across thirteen convolutional layers, thereby enabling the extraction of low- and mid-level textural features [31].

ResNet50V2 is an improved derivative of the original ResNet50 architecture, featuring pre-activated residual blocks that implement batch normalization and ReLU activation before convolution operations. This structure improves gradient flow in deep networks and supports feature reuse between layers, resulting in more stable convergence behavior and stronger high-level feature representations [32].

MobileNetV2 performs computationally efficient feature extraction with far fewer parameters compared to VGG16 and ResNet50V2 by using depth-decomposable convolutions combined with inverse residual bottleneck structures. The architecture is particularly successful in capturing finely detailed spatial features while maintaining low computational cost [33].

Rather than guessing at a freezing strategy, we tested one. Under identical training conditions, we compared seven setups—from backbones that were completely frozen (with only the newly added classification head learning anything) through to backbones left fully open, where every layer of VGG16, ResNet50V2, and MobileNetV2 was fine-tuned. The fully frozen version clearly underperformed the others, which tells us the backbones do need to adapt to lung CT data rather than relying on ImageNet features as-is. Among what remained, fine-tuning all three networks end-to-end gave the best results and held up most consistently across the five folds, so that is what we kept for the final model.

All three models are initialized with weights pre-trained on ImageNet and have a common input size of $224 \times 224 \times 3$. The classification layers of each model are replaced with a Global Average Pooling layer that reduces the output feature maps to fixed-length vectors. The three resulting vectors are combined to form an integrated model, which passes through a fully connected layer of 512 neurons with ReLU activation, a 50% Dropout regularization, and finally a three-neuron output layer with softmax activation that generates class probabilities. All architectural components of the proposed ensemble model are as shown in Figure 1.

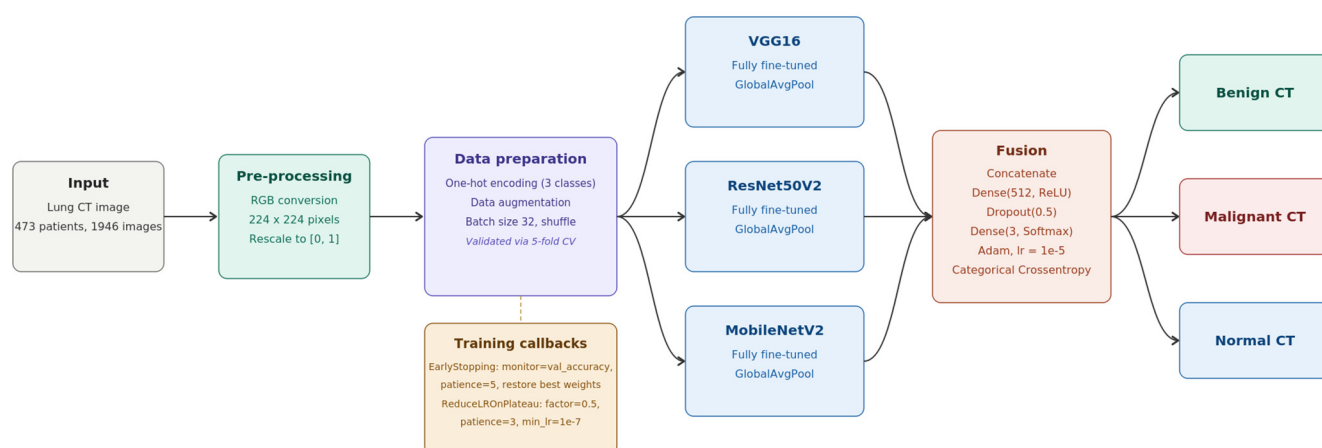


Figure 1. All components of the proposed model.

3.5. Model Training and Optimization

The ensemble model was compiled with the Adam optimizer, which has an initial learning rate of 1×10^{-5} . Adam was chosen because it can perform independent learning rate updates for each parameter; this feature significantly stabilizes the training process in medical imaging data. Categorical cross-entropy was used as the loss function for multi-class classification:

$$L(y, \hat{y}) = -\sum_i y_i \log(\hat{y}_i)$$

Model performance was monitored at each epoch separately for both the training and validation subsets using accuracy, precision, and sensitivity metrics. Training is limited to a maximum of 30 epochs with a batch size of 32.

Two callbacks were used to prevent overfitting and make training more efficient. Early Stopping was configured to halt training and restore the best weights if validation accuracy did not improve over 5 consecutive epochs. ReduceLROnPlateau, on the other hand, halves the learning rate if the validation loss did not decrease over three epochs, allowing the model to converge more precisely.

The training parameters of the proposed model are given in Table 2.

Table 2. Training parameters of the proposed model.

Parameter	Value
Batch Size	32
Maximum Number of Epochs	30
Initial Learning Rate	1×10^{-5}
Optimizer	Adam
Loss Function	categorical_crossentropy
Cross-Validation Procedure	5-fold GroupKFold (patient-level)
Training/Validation Separation	80%/20% (patient-level)
Early Stopping patience duration (patience)	5 epochs (val_accuracy, max mode)
ReduceLROnPlateau multiplier	0.5 (patience = 3 epochs)
Minimum Learning Rate	1×10^{-7}

3.6. Performance Metrics

Model performance was evaluated across six complementary metrics. Accuracy gives the percentage of correctly classified samples among all samples. Precision shows how well the samples the model identifies as positive match those that are actually positive, while sensitivity measures how many true positives were detected. The F1 score combines these

two metrics into a single value, providing a more reliable summary, especially in cases of unbalanced class distributions.

True Positive (TP): These are cases where the model made a correct prediction.

- Benign: There is indeed a benign nodule in the image, and the model recognized it as benign.
- Malignant: There is indeed a malignant nodule in the image, and the model recognized it as malignant.
- Normal: There are no pathological findings in the image, and the model correctly identified this as normal.

False Positive (FP): These are situations where the model makes something that does not exist appear to exist.

- Benign: Although no benign nodule was detected in the image, the model classified it as benign.
- Malignant: Although no malignant nodule was detected in the image, the model classified it as malignant.
- Normal: Although a nodule was present in the image, the model evaluated it as normal.

True Negative (TN): These are cases where the model correctly identifies that an image does not belong to a particular class.

- Benign: The image shows no benign nodules, and the model accurately distinguishes between malignant and normal nodules.
- Malignant: The image shows no malignant nodules, and the model correctly distinguishes it as either benign or normal.
- Normal: A nodule is present in the image, and the model has correctly classified it as either benign or malignant.

False Negative (FN): This is the most clinically risky type of error, where the model misses something that already exists.

- Benign: The image shows a benign nodule, but the model has misclassified it as either malignant or normal.
- Malignant: The image shows a malignant nodule, but the model has misclassified it as benign or normal. This type of error is particularly dangerous; missing a malignant case can lead to delayed treatment.
- Normal: Despite the absence of any pathological findings in the image, the model incorrectly assessed it as either benign or malignant.

Accuracy is a fundamental metric that measures the overall success of a model. It is defined as the percentage of correctly classified items among all samples and serves as a quick indicator of performance in balanced datasets.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

Precision measures the accuracy of positive predictions. It expresses the percentage of examples that are actually positive among those that the model qualifies as positive; high precision corresponds to a low false positive rate.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

Recall/Sensitivity indicates the extent to which the model can capture all true positives in the dataset. It is defined as the ratio of true positives to the total number of true positives; high sensitivity indicates a low false negative rate.

$$\text{Recall/Sensitivity} = \text{TP} / (\text{TP} + \text{FN})$$

The F1 score is the harmonic mean of precision and recall. It offers a more reliable measure than the accuracy metric when class distribution is uneven, as it takes into account both false positives and false negatives.

$$\text{F1} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

The ROC curve and AUC visually demonstrate the model's ability to discriminate between classes. The AUC value numerically expresses the model's capacity for accurate classification; the higher the AUC value, the more successful the model.

$$\text{AUC} \approx \sum_i (\text{FPR}_{i+1} - \text{FPR}_i) \cdot (\text{TPR}_{i+1} + \text{TPR}_i) / 2$$

The Confusion Matrix is a table that details the performance of a classification model on a test set where the actual class labels are known. Each row of the matrix represents the predicted class, and each column represents the actual class; this structure allows for a visual examination of the algorithm's error patterns.

3.7. Model Interpretability

To check whether the ensemble is actually looking at clinically meaningful regions when it calls a case malignant, we applied Grad-CAM to the final convolutional layer of each backbone (VGG16: block5_conv3; ResNet50V2: post_relu; MobileNetV2: out_relu) and averaged the three activation maps into a single fused attention map. We focus this analysis on malignant cases, since correctly localizing malignancy carries the greatest clinical weight.

4. Results

Before training the final model, the stability of the selected configuration was assessed using five-fold, patient-level GroupKFold cross-validation over the full dataset. Table 3 reports the accuracy and macro F1 score obtained in each fold, and Table 4 reports the corresponding class-wise precision, recall, and F1 score, averaged across the five folds.

Table 3. Five-fold cross-validation results by fold.

Fold	Accuracy	Macro F1	Epochs
1	0.8769	0.8437	12
2	0.9049	0.8840	11
3	0.9460	0.9178	22
4	0.8715	0.8529	23
5	0.8715	0.8275	14
Mean $\pm$ SD	0.8942 $\pm$ 0.0321	0.8652 $\pm$ 0.0359	16.4 $\pm$ 5.7

Table 4. Class-wise precision, recall, and F1 score (mean $\pm$ SD across five folds).

Class	Precision	Recall	F1 Score
Benign	0.7682 $\pm$ 0.1063	0.7223 $\pm$ 0.0875	0.7412 $\pm$ 0.0743
Malignant	0.9284 $\pm$ 0.0487	0.8892 $\pm$ 0.0530	0.9069 $\pm$ 0.0287
Normal	0.9163 $\pm$ 0.0394	0.9816 $\pm$ 0.0294	0.9475 $\pm$ 0.0283

The variation across folds is fairly small (± 3.21 percentage points in accuracy), which tells us performance here is not riding on one lucky split—it holds up across different groups of patients. Benign is the exception: precision swings the most from fold to fold (0.7682 ± 0.1063), which fits with there being fewer benign patients in the dataset and the visual overlap between benign nodules and normal tissue we discuss in Section 5. This cross-validation was about confirming the configuration is stable, not about producing the headline number—the figures reported for the rest of this section come from retraining the same setup on a separate, patient-level 80/20 split, as described in Section 3.2. The resulting class-based classification report on the validation set is presented in Table 5.

Table 5. Class-based classification report in the validation set.

Class	Precision	Recall	F1 Score
Benign	0.8596	0.6901	0.7656
Malignant	0.8750	0.9545	0.9130
Normal	0.9395	0.9484	0.9439
Accuracy			0.9062
Macro Avg	0.8914	0.8643	0.8742
Weighted Avg	0.9054	0.9062	0.9037

Broken down by class, the model reached 85.96% precision and 69.01% recall (F1: 0.7656) on benign samples, 87.50% precision and 95.45% recall (F1: 0.9130) on malignant samples, and 93.95% precision and 94.84% recall (F1: 0.9439) on normal samples. Overall accuracy came to 90.62%, with macro-averaged precision, recall, and F1 of 89.14%, 86.43%, and 0.8742.

The three backbones clearly complement each other here, but performance is not the same across classes. Malignant cases stand out, with recall reaching 95.45%; the clinical significance of this figure is discussed in Section 5. Benign cases are the weaker spot: the lower recall there suggests the model still confuses some benign nodules with normal tissue. The corresponding accuracy and loss curves are as shown in Figure 2.

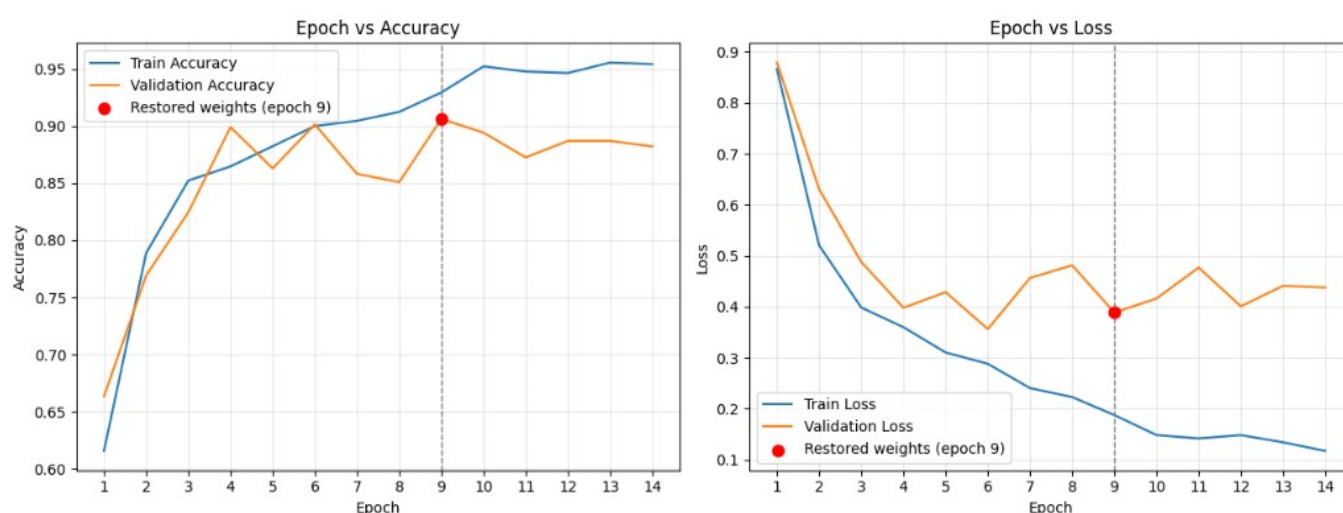


Figure 2. Accuracy and loss graphs.

The corresponding confusion matrix and ROC curves are as shown in Figure 3.

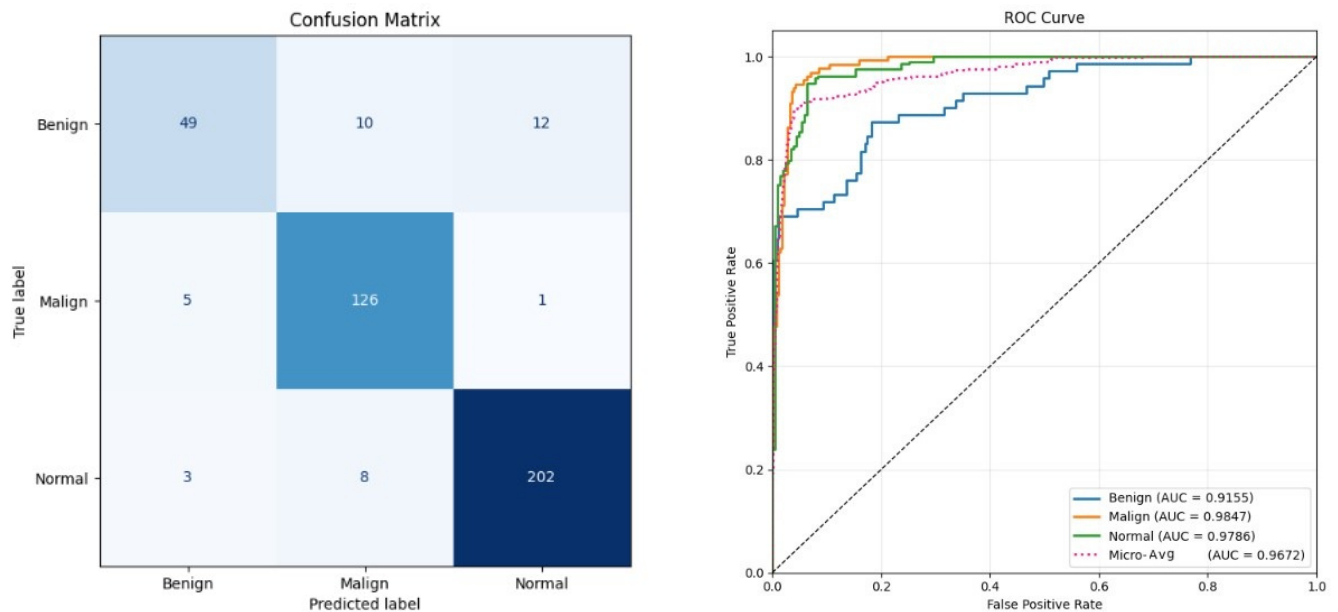


Figure 3. Confusion matrix and ROC curve graphs.

According to the confusion matrix, 49 of 71 benign samples, 126 of 132 malignant samples, and 202 of 213 normal samples were correctly classified, resulting in a total of 39 misclassifications across the validation set. Receiver Operating Characteristic (ROC) analysis yielded an Area Under the Curve (AUC) of 0.9847 for the malignant class and 0.9786 for the normal class, alongside a micro-averaged AUC of 0.9672 across all three classes, demonstrating strong overall discriminative capability. The benign class showed a lower AUC of 0.9155, which aligns with its lower precision and recall relative to the other categories.

Errors within the benign class were distributed roughly evenly between the malignant and normal classes (10 and 12 samples misclassified, respectively) rather than skewed toward a single direction. This pattern is consistent with the literature and likely reflects the morphological overlap between benign nodules and both malignant lesions and normal parenchyma at the current imaging resolution. By contrast, malignant cases were classified with high recall; only 6 of 132 samples were misclassified (5 as benign, 1 as normal), reflecting the model's strong sensitivity to malignant tissue patterns. Of the 213 normal samples, 11 were misclassified (8 as malignant, 3 as benign); the higher rate of normal-to-malignant confusion slightly increases the false-positive rate for this class, but this trade-off does not carry the critical clinical risk associated with missing an actual malignant case.

During training, the loss decreased consistently across epochs, and the validation loss stabilized following the learning rate reductions triggered by the ReduceLROnPlateau callback, indicating that the optimization configuration effectively managed network convergence.

To empirically validate the proposed ensemble architecture, we evaluated each backbone on its own before combining them into the full ensemble. The results are in Table 6.

The ensemble beats every individual backbone on both metrics. MobileNetV2 stands out here—it performed noticeably worse than the others on its own (F1: 0.5555). We kept it anyway, rather than swapping in VGG19 or ResNet50, even though both scored higher alone (F1: 0.8335 and 0.8006). The reasoning: VGG19 and ResNet50 are close relatives of VGG16 and ResNet50V2, which are already in the ensemble, so they would likely make the same kinds of mistakes rather than different ones. MobileNetV2 works differently—it is built on depthwise separable convolutions—so we picked it for that architectural diversity, not for its solo accuracy. And that choice paid off: its presence did not stop the ensemble

from reaching the best overall result, which fits with the idea that what drives ensemble performance is diversity across architectures, not how strong any one piece is on its own.

Table 6. Ablation study results—individual backbone performance.

Backbone	Accuracy	F1 Score
VGG16	0.8670	0.8356
VGG19	0.8646	0.8335
ResNet50	0.8389	0.8006
ResNet50V2	0.8598	0.8284
MobileNetV2	0.7572	0.5555
EfficientNetB0	0.5385	0.3218
VGG16 + ResNet50V2 + MobileNetV2 (Proposed model)	0.9062	0.8742

Freezing strategy was not something we assumed—we tested it. Under identical training conditions, we ran five-fold cross-validation on seven setups: fully frozen backbones, three intermediate points (25%, 50%, 75% of layers trainable), the layer-specific scheme originally considered (VGG16: first 15 layers, ResNet50V2: first 50, MobileNetV2: first 100), a two-phase warm-up variant, and fully open fine-tuning. Table 7 has the full comparison.

Table 7. Comparison of freezing strategies under five-fold cross-validation (mean $\pm$ SD).

Configuration	Accuracy (%)	Macro F1	Benign F1	Malignant F1	Normal F1
Fully frozen (0% trainable)	84.94 $\pm$ 3.76	0.798 $\pm$ 0.048	0.597 $\pm$ 0.093	0.855 $\pm$ 0.040	0.942 $\pm$ 0.028
25% trainable	88.64 $\pm$ 2.75	0.852 $\pm$ 0.030	0.707 $\pm$ 0.053	0.906 $\pm$ 0.028	0.942 $\pm$ 0.024
50% trainable	88.59 $\pm$ 3.67	0.855 $\pm$ 0.039	0.720 $\pm$ 0.059	0.903 $\pm$ 0.033	0.942 $\pm$ 0.029
75% trainable	87.20 $\pm$ 3.69	0.837 $\pm$ 0.037	0.689 $\pm$ 0.046	0.884 $\pm$ 0.040	0.939 $\pm$ 0.032
Layer-specific scheme (15/50/100)	87.36 $\pm$ 2.16	0.837 $\pm$ 0.021	0.687 $\pm$ 0.043	0.882 $\pm$ 0.027	0.943 $\pm$ 0.022
Two-phase (warm-up + partial)	87.26 $\pm$ 2.85	0.840 $\pm$ 0.029	0.700 $\pm$ 0.037	0.881 $\pm$ 0.030	0.940 $\pm$ 0.026
Fully open (100% trainable) (Adopted)	89.42 $\pm$ 3.21	0.865 $\pm$ 0.036	0.741 $\pm$ 0.074	0.907 $\pm$ 0.029	0.948 $\pm$ 0.028

Freezing everything performed clearly worse than every other option, so some adaptation to lung CT data is genuinely needed. Beyond that, the differences were fairly small, but fully fine-tuning all three backbones came out on top, with the highest accuracy and macro F1, and it held up consistently across folds. That is what we kept for the final model.

To address the concern regarding cross-dataset comparability, the ensemble configuration proposed by Kumaran et al. [29] was trained and evaluated on our dataset using the training configuration described in their own study, so that the architecture was assessed under the settings its authors intended rather than under settings tuned for our model. Since patient identifiers are available for our institutional dataset, we applied the same patient-level, leakage-free split used throughout this study to the Kumaran et al. [29] replication as well, rather than following their original image-level protocol. This ensures a fully matched comparison in terms of data partitioning. Table 8 reports the results of this cross-dataset comparison with Kumaran et al. [29].

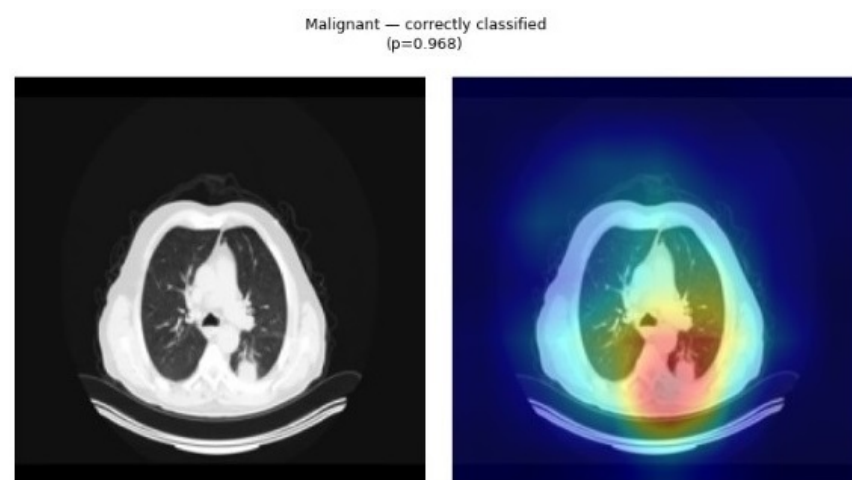
Table 8. Cross-dataset comparison with Kumaran et al. [29].

Study	Model	Dataset	Accuracy	Macro F1
Kumaran et al. [29] (reported)	VGG16+ResNet50 +InceptionV3	IQ-OTH/NCCD	98.18%	0.969
Kumaran et al. [29]	VGG16+ResNet50 +InceptionV3	Our dataset (Their own criteria) (Patient-level)	80.77%	0.758
Proposed model	VGG16+ResNet50V2 +MobileNetV2	IQ-OTH/NCCD (Our criteria)	96.82%	0.930
Proposed model	VGG16+ResNet50V2 +MobileNetV2	Our dataset (Patient-level)	90.62%	0.874

The configuration proposed by Kumaran et al. [29] showed a substantial decline in performance when transferred to our dataset (98.18% to 80.77% accuracy), whereas the proposed model's performance on IQ-OTH/NCCD remained comparable to, and in fact slightly exceeded, its performance on our own dataset (90.62% to 96.82% accuracy). This asymmetry suggests that the proposed architecture generalizes more robustly across different imaging sources than the compared configuration, and that its reported performance is not confined to a single dataset.

One methodological point should be noted regarding this comparison. Since patient identifiers are available for our institutional dataset, the same patient-level, leakage-free split used throughout this study was also applied to the Kumaran et al. [29] replication, rather than the image-level protocol used in their original work. This ensures that the two models are compared under an identical, matched data-partitioning scheme, so the reported gap (90.62% vs. 80.77%) cannot be attributed to any difference in splitting strategy between the two configurations. A separate caveat applies only to the IQ-OTH/NCCD control experiment: as this public dataset does not provide patient identifiers, an image-level split was necessarily used there. The 96.82% accuracy reported for that experiment should therefore be interpreted with the possibility of residual image-level leakage in mind, unlike all other results in this study.

Figure 4 shows a representative Grad-CAM visualization for a malignant case, illustrating the image regions contributing most to the model's malignant classification decision.

**Figure 4.** Grad-CAM visualization for a representative malignant case.

To get a sense of where the model is actually looking when it flags a case as malignant, Grad-CAM was applied to a representative example. The highlighted region lines up closely

with the perihilar/mediastinal area containing the lesion, consistent with the model's strong overall recall for malignant cases (95.45%).

5. Discussion

The results show that ensemble transfer learning is a genuinely competitive approach for classifying lung cancer from CT images. Table 9 places our results alongside those reported in the literature, though, as discussed below, these comparisons come with real caveats.

Table 9. Summary of literature studies on lung cancer classification from CT images.

Study	Dataset	Method	Accuracy
Çiftçi (2022) [12]	TCIA	MobileNetV3	81%
		VGG16	84%
		InceptionV4	86%
		AlexNet	96%
Üzülmez and Çiftçi (2024) [13]	LUNA16	4-layer CNN	97.1%
Şeker et al. (2024) [14]	C25000 (Kaggle) Histopathology	CNN	87.93%
		ViT	98.40%
		EfficientNet-B1	96.97%
Shah et al. (2023) [15]	LUNA16	Ensemble 2D CNN for nodule detection.	95.00%
Wankhade and Vigneshwari (2023) [17]	LIDC-IDRI	CCDC-HNN: 3D-CNN hybrid	95.00%
Wani et al. (2024) [19]	Survey Lung Cancer (Kaggle)	DeepXplainer: CNN + XGBoost	97.43%
Kusuma et al. (2024) [24]	IQ-OTH/NCCD	Hybrid CNN-RNN + Pelican Optimization	97.30%
Reshma et al. (2024) [25]	IQ-OTH/NCCD (Mendeley)	Deep CNN	95.00%
Mohana Krishna and Puviarasi (2024) [26]	IQ-OTH/NCCD	ResNet50 vs. InceptionV3	93.09%
Safta and Shaffie (2024) [27]	LIDC-IDRI	3D-LOP + 3D-CNN	97.84%
Kumaran et al. (2024) [29]	IQ-OTH/NCCD	VGG16 + ResNet50 + InceptionV3	98.18%
Proposed Model	IQ-OTH/NCCD	VGG16 + ResNet50V2 + MobileNetV2	96.82%
Proposed Model	Own dataset (Patient Level)	VGG16 + ResNet50V2 + MobileNetV2	90.62%

Direct numerical comparisons across studies in Table 9 should be interpreted with caution, as each study was evaluated on a different dataset. Performance differences likely reflect variations in dataset size, diversity, and clinical characteristics, not only architectural choices.

Most referenced studies used public benchmarks such as LUNA16, LIDC-IDRI, or IQ-OTH/NCCD. In contrast, this study was conducted on CT images from 473 patients collected at Atatürk University Hospital. Results obtained on such institutional data may better reflect local clinical conditions and patient demographics.

To establish a controlled baseline within our dataset, we conducted an ablation study in which each backbone model was trained and evaluated independently (Table 6), and a direct re-implementation of the most closely related ensemble configuration, that of Kumaran et al. [29], trained on our dataset under the settings reported in their own study (Table 8). This configuration showed a substantial drop in performance when

evaluated on our institutional data (80.77% accuracy, macro F1 0.758) compared to the 98.18% originally reported, reinforcing the view that the accuracy figures in Table 9 are not directly comparable across studies.

The ensemble's strong performance can largely be explained by how differently its three backbones represent the input. ResNet50V2's pre-activated residual blocks give it a more stable gradient flow during fine-tuning than a standard ResNet50 would offer, which in turn helps it pick up higher-level features tied to lung parenchyma. MobileNetV2 works differently: its inverted residual bottlenecks let it stay parameter-efficient while still picking up fine spatial detail. Together with VGG16's tissue-sensitive deep features and the skip-connection-rich representations from ResNet50V2, the three networks end up covering somewhat different parts of the feature space. For the classifier head, we tested several configurations and settled on a single 512-neuron Dense layer; this was mostly an empirical decision, and combined with 50% Dropout, it simply gave more consistent validation results than the deeper heads we tried.

A common shortcoming in ensemble studies in this area is reporting only the final combined accuracy, without checking whether each backbone is actually contributing. We tried to avoid that here: the freezing-strategy ablation in Section 4 (Table 7) and the per-backbone breakdown in Table 6 both indicate that the three models are contributing in genuinely different ways, not simply combined for convenience.

Clinically, the recall on malignant cases is probably the number that matters most. Mistaking a malignant case for a benign one is a much costlier error here than the reverse, since it can delay diagnosis and shrink the window for effective treatment. The model reached 95.45% recall for the malignant class, meaning the vast majority of true malignant cases were correctly caught, exactly the kind of error the system needs to keep to a minimum.

Benign cases were where the model struggled most. That is not entirely surprising: some benign lesions look a lot like normal parenchymal variation at the resolution typical CT imaging offers, and this overlap shows up throughout the literature. It probably did not help that benign was also our smallest class (356 of 1946 images, 18.3%), largely because relatively few benign-nodule patients had scans with nodules clear enough to include in the dataset in the first place.

Interestingly, the benign misclassifications were not skewed toward one direction: errors split fairly evenly between malignant (10 cases) and normal (12 cases), which fits with the morphological-overlap explanation above rather than pointing to some systematic bias.

This study did not separately annotate nodule boundaries, so a size-based analysis was not possible. That said, small nodules and low contrast-to-noise ratios are a known source of misclassification in CT-based analysis, so some of the benign errors we see may stem from this; it is a natural thing to test once nodule-level size annotations are available.

Two paths seem worth pursuing in the meantime. Spatial attention, which weights small, low-contrast regions more heavily, could help the model pick up nodules it currently misses. Multi-scale feature fusion is a related option, since fine detail lost in coarser feature maps is often what a small nodule needs to be recognized. On the clinical side, a cost-sensitive decision threshold at inference time, rather than retraining, offers a way to trade off benign recall against malignant recall as needed.

6. Conclusions

This study set out to build a multi-backbone ensemble for classifying lung cancer from CT images, bringing together VGG16, ResNet50V2, and MobileNetV2. Tested on 1946 CT images from 473 patients, the model reached 90.62% accuracy with a macro F1 score of 0.874. Combining the three backbones gave a modest but consistent edge over any single

model on its own, and this was not just assumed: it was checked empirically through cross-validation and ablation.

Data augmentation, ReduceLROnPlateau, and EarlyStopping together kept training stable without needing synthetic sampling methods like SMOTE. The model's recall on malignant cases, 95.45%, means it rarely misses an actual cancer, which is the error that matters most clinically. Benign cases were harder: they are both underrepresented in the dataset and morphologically similar to normal tissue, and closing that gap is a natural next step. Spatial attention and a cost-sensitive inference threshold, discussed in Section 5, offer concrete starting points for that work.

The ablation work backs up the design choices throughout. Each backbone on its own performed noticeably worse than the full ensemble (Table 6), and a systematic comparison of freezing strategies confirmed that fully fine-tuning all three backbones beat both frozen and partially frozen setups (Table 7). Reproducing Kumaran et al.'s [29] closely related ensemble under the settings reported in their own study showed that their reported numbers did not carry over to our institutional data, which says something important about comparing across datasets rather than within one.

The proposed architecture was also applied to IQ-OTH/NCCD, an existing public dataset, and achieved strong results there as well (96.82% accuracy). The current framework relies solely on CT images, though; folding in complementary modalities such as PET/CT, which already informed the ground-truth labeling in this study, may further improve diagnostic accuracy, particularly for the benign class where morphological overlap with normal tissue was most pronounced.

As the model was developed and evaluated on retrospective, single-center data, claims of clinical applicability should be regarded as preliminary until confirmed through multicenter validation and prospective clinical studies.

Future work should prioritize multicenter validation across different scanners and institutions, exploration of more advanced architectures, extension of the explainability analysis introduced here through additional techniques, and prospective clinical evaluation to establish real-world diagnostic utility.

Author Contributions: Conceptualization, İ.M.S.; Methodology, İ.M.S. and M.Y.; Software, İ.M.S.; Validation, İ.M.S., M.Y. and A.M.; Formal analysis, İ.M.S. and M.Y.; Investigation, İ.M.S. and A.M.; Resources, A.M.; Data curation, İ.M.S. and A.M.; Writing—original draft, İ.M.S.; Writing—review & editing, İ.M.S. and M.Y.; Visualization, İ.M.S.; Supervision, M.Y.; Project administration, İ.M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by Ataturk University Scientific Research Projects Coordination Unit. Project number: FDK-2022-11341.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Institutional Ethics Committee of Ataturk University (2025-1/21 numbered. Date of Approval: 31 January 2025).

Informed Consent Statement: Patient consent was waived due to the retrospective nature of the study and the use of fully anonymized archival data.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Karakurt, M.; İşeri, İ. Patoloji görüntülerinin derin öğrenme yöntemleri ile sınıflandırılması. *Avrupa Bilim Teknol. Derg.* **2022**, *33*, 192–206. [\[CrossRef\]](#)
- Agarwal, S.; Yadav, A.S.; Dinesh, V.; Vatsav, K.S.S.; Prakash, K.S.S.; Jaiswal, S. By artificial intelligence algorithms and machine learning models to diagnosis cancer. *Mater. Today Proc.* **2023**, *80*, 2969–2975. [\[CrossRef\]](#)
- Gao, Q.; Yang, L.; Lu, M.; Jin, R.; Ye, H.; Ma, T. The artificial intelligence and machine learning in lung cancer immunotherapy. *J. Hematol. Oncol.* **2023**, *16*, 55. [\[CrossRef\]](#) [\[PubMed\]](#)
- Espinoza, J.L.; Dong, L.T. Artificial intelligence tools for refining lung cancer screening. *J. Clin. Med.* **2020**, *9*, 3860. [\[CrossRef\]](#) [\[PubMed\]](#)
- Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; Van Der Laak, J.A.; Van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [\[CrossRef\]](#) [\[PubMed\]](#)
- Mohammed, M.A.; Mukhlif, A.A.; Khateeb, B.A. Breast cancer images classification using a new transfer learning technique. *Iraqi J. Comput. Sci. Math.* **2023**, *4*, 16. [\[CrossRef\]](#)
- Altuhaifa, F.A.; Win, K.T.; Su, G. Predicting lung cancer survival based on clinical data using machine learning: A review. *Comput. Biol. Med.* **2023**, *165*, 107338. [\[CrossRef\]](#) [\[PubMed\]](#)
- Hussain, L.; Almaraashi, M.S.; Aziz, W.; Habib, N.; Saif Abbasi, S.-U.-R. Machine learning-based lungs cancer detection using reconstruction independent component analysis and sparse filter features. *Waves Random Complex Media* **2024**, *34*, 226–251.
- Alex, S.A.; Nayahi, J.J.V.; Kaddoura, S. Deep convolutional neural networks with genetic algorithm-based synthetic minority over-sampling technique for improved imbalanced data classification. *Appl. Soft Comput.* **2024**, *156*, 111491. [\[CrossRef\]](#)
- Thakur, A.; Gupta, M.; Sinha, D.K.; Mishra, K.K.; Venkatesan, V.K.; Guluwadi, S. Transformative breast Cancer diagnosis using CNNs with optimized ReduceLROnPlateau and Early stopping Enhancements. *Int. J. Comput. Intell. Syst.* **2024**, *17*, 14. [\[CrossRef\]](#)
- Albalawi, E.; Mahesh, T.R.; Thakur, A.; Kumar, V.V.; Gupta, M.; Khan, S.B.; Almusharraf, A. Integrated approach of federated learning with transfer learning for classification and diagnosis of brain tumor. *BMC Med. Imaging* **2024**, *24*, 110. [\[CrossRef\]](#) [\[PubMed\]](#)
- Cifci, M. Derin öğrenme metodu kullanarak BT görüntülerinden akciğer kanseri teşhisi. *Dokuz Eylül Üniversitesi Mühendis. Fakültesi Fen. Mühendis. Derg.* **2022**, *24*, 487–500. [\[CrossRef\]](#)
- Üzülmez, S.; Cifci, M.A. Early diagnosis of lung cancer using deep learning and uncertainty measures. *J. Fac. Eng. Archit. Gazi Univ.* **2024**, *39*, 385–400.
- Şeker, D.; Kartal, M.S.; Yıldız, A.; Öksüz, İ. Akciğer Kanseri Tespitinde Dönüşüm ve Evrişim Tabanlı Modeller ile Açıklanabilir Yapay Zeka Uygulaması. *EMO Bilimsel Dergi* **2024**, *14*, 59–69.
- Shah, A.A.; Malik, H.A.M.; Muhammad, A.; Alourani, A.; Butt, Z.A. Deep learning ensemble 2D CNN approach towards the detection of lung cancer. *Sci. Rep.* **2023**, *13*, 2987. [\[CrossRef\]](#) [\[PubMed\]](#)
- Mikhael, P.G.; Wohllwend, J.; Yala, A.; Karstens, L.; Xiang, J.; Takigami, A.K.; Bourgooin, P.P.; Chan, P.; Mrah, S.; Amayri, W. Sybil: A validated deep learning model to predict future lung cancer risk from a single low-dose chest computed tomography. *J. Clin. Oncol.* **2023**, *41*, 2191–2200. [\[CrossRef\]](#) [\[PubMed\]](#)
- Wankhade, S.; Vigneshwari, S. A novel hybrid deep learning method for early detection of lung cancer using neural networks. *Healthc. Anal.* **2023**, *3*, 100195. [\[CrossRef\]](#)
- Said, Y.; Alsheikhy, A.A.; Shawly, T.; Lahza, H. Medical images segmentation for lung cancer diagnosis based on deep learning architectures. *Diagnostics* **2023**, *13*, 546. [\[CrossRef\]](#) [\[PubMed\]](#)
- Wani, N.A.; Kumar, R.; Bedi, J. DeepXplainer: An interpretable deep learning based approach for lung cancer detection using explainable artificial intelligence. *Comput. Methods Programs Biomed.* **2024**, *243*, 107879. [\[CrossRef\]](#) [\[PubMed\]](#)
- Chae, K.J.; Lim, S.; Seo, J.B.; Hwang, H.J.; Choi, H.; Lynch, D.; Jin, G.Y. Interstitial lung abnormalities at CT in the Korean National Lung Cancer Screening Program: Prevalence and deep learning-based texture analysis. *Radiology* **2023**, *307*, e222828. [\[CrossRef\]](#) [\[PubMed\]](#)
- Guan, P.; Yu, K.; Wei, W.; Tan, Y.; Wu, J. Big data analytics on lung cancer diagnosis framework with deep learning. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2023**, *21*, 757–768. [\[CrossRef\]](#) [\[PubMed\]](#)
- Mohamed, T.I.; Oyelade, O.N.; Ezugwu, A.E. Automatic detection and classification of lung cancer CT scans based on deep learning and ebola optimization search algorithm. *PLoS ONE* **2023**, *18*, e0285796. [\[CrossRef\]](#) [\[PubMed\]](#)
- Rajasekar, V.; Vaishnave, M.; Premkumar, S.; Sarveshwaran, V.; Rangaraaj, V. Lung cancer disease prediction with CT scan and histopathological images feature analysis using deep learning techniques. *Results Eng.* **2023**, *18*, 101111. [\[CrossRef\]](#)
- Kusuma, S.; Krishnan, S.G.; Samreen, K.; Ramana, M.V.; Prasad, G.S. A hybrid deep learning approach for early detection and classification of lung cancer using the pelican optimization algorithm. In Proceedings of the 2024 International Conference on Distributed Computing and Optimization Techniques (ICDCOT), Bengaluru, India, 15–16 March 2024; pp. 1–6.

25. Reshma, G.; Jeyavathana, R.; Suguna, M. Performance analysis of lung cancer detection and its classification using deep convolutional neural network with low dose Ct scan lung images over convolutional neural network with increase in classifier accuracy. *AIP Conf. Proc.* **2024**, 2729, 060024. [[CrossRef](#)]
26. Mohana Krishna, N.; Puviarasi, R. Convolutional neural network based ResNet50 for finding accuracy in prediction of lung cancer using CT images and compared with CNN based inception V3. *AIP Conf. Proc.* **2024**, 2816, 200002. [[CrossRef](#)]
27. Safta, W.; Shaffie, A. Advancing pulmonary nodule diagnosis by integrating Engineered and Deep features extracted from CT scans. *Algorithms* **2024**, 17, 161. [[CrossRef](#)]
28. Princy Magdaline, P.; Ganesh Babu, T. Detection of lung cancer using novel attention gate residual U-Net model and KNN classifier from computer tomography images. *J. Intell. Fuzzy Syst.* **2023**, 45, 6289–6302. [[CrossRef](#)]
29. Kumaran, S.Y.; Jeya, J.J.; Mahesh, T.R.; Khan, S.B.; Alzahrani, S.; Alojail, M. Explainable lung cancer classification with ensemble transfer learning of VGG16, Resnet50 and InceptionV3 using grad-cam. *BMC Med. Imaging* **2024**, 24, 176. [[CrossRef](#)] [[PubMed](#)]
30. Yang, X.; Duan, A.; Jiang, Z.; Li, X.; Wang, C.; Wang, J.; Zhou, J. Segmentation and Classification of Lung Cancer Images Using Deep Learning. *Appl. Sci.* **2026**, 16, 628. [[CrossRef](#)]
31. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 630–645.
33. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–28 June 2018; pp. 4510–4520.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.