

Automated Diagnosis of Lung Cancer Subtypes Using Deep Convolutional Neural Network

Manash Sarker, Md Intishar Ahmed*, Sadia Zannat Reem

Department of Computer and Communication Engineering, PSTU

Correspondence: intishar17@cse.pstu.ac.bd

(Received Date: 4th May 2026, Accepted Date: 20th September 2026)

Abstract: Lung cancer subtype classification based on CT scan images is difficult to achieve owing to the similar visual appearance of malignant subtypes. In this research, transfer learning-based CNNs are examined for four class classification of lung CT images into adenocarcinoma, large cell carcinoma, squamous cell carcinoma, and normal lung CT images. Transfer learning-based architectures such as ResNet-50, VGG-16, and Xception are evaluated in which ResNet-50 is chosen as the main architecture based on the overall classification results. To ensure a more reliable assessment of the classification performance beyond a single train-validation split, five-fold cross-validation method is used with 1,000 images in which 800 images are used for training and 200 images are used for validation in each fold. The ResNet-50 model achieved a mean accuracy of $92.70 \pm 2.46\%$, macro precision of $93.69 \pm 2.51\%$, macro recall of $92.48 \pm 2.06\%$, and macro F1-score of $92.95 \pm 2.27\%$. The out-of-fold confusion matrix revealed that the classification of normal lung images is achieved with very high accuracy while most of the errors belong to malignant subtypes including adenocarcinoma, large-cell carcinoma, and squamous-cell carcinoma. These results indicate that transfer learning can provide strong computational performance for multiclass lung CT image classification under the evaluated experimental conditions. However, patient-level validation and external evaluation on independently collected clinical datasets are required before clinical applicability can be established.

Keywords: CNN, CT scan imaging, deep learning, Lung cancer classification, ResNet-50, VGG-16, Xception.

Introduction

Lung cancer is one of the major causes of morbidity and mortality associated with cancer diseases across the world. As per Global Cancer Statistics 2020, an estimation of 19.3 million new cases and nearly 10 million cancer deaths were recorded across the world in 2020 (Chhikara & Parang, 2023). Of the total cancer diagnoses, 2.21 million cases and 1.79 million deaths were due to lung cancer, making it one of the major reasons for mortality among cancer cases (Sung et al., 2021). An accurate and timely diagnosis is crucial for proper management of the disease. However, visual interpretation of images by radiologists may prove difficult because of variations in appearances of lesions. Computer-Aided Diagnosis (CAD) approaches have hence been explored for assisting in analyzing images from the thoracic region. CT is particularly useful in diagnosing lung cancer as it offers anatomical information and allows for visualization of abnormalities in the lungs (Tekade, 2018).

Types of lung cancer have been classified into small cell lung cancer (SCLC) and non-small cell lung cancer (NSCLC), where NSCLC accounts for most of the types of lung cancers (Herbst et al., 2008). Main types of NSCLC include adenocarcinoma, squamous cell carcinoma and large cell carcinoma. Classification of these types is important considering the fact that their histological features differ in terms of behavior, prognosis and treatment options. This has made automated classification of NSCLC from CT images an active area of research. Lung cancer detection, which involves detecting the presence of a tumor, and lung cancer classification, which can classify diagnostic groups, should be differentiated from histological subtype classification.

The application of deep learning, especially the CNN, has shown promise in the automatic analysis of lung CT images. Initial studies analyzed traditional machine learning methods, as well as CNN, in terms of the ability to classify lung nodules and cancers. Further, VGG, DenseNet and ResNet architectures were used to analyze automated thoracic image analysis (Tekade, 2018; Ausawalaithong et al., 2018; Pandian et al., 2022; Nazir et al., 2023; Muntasir et al., 2023). Recent research has focused on transfer learning, attention mechanisms, radiomic features and the comparison of diverse deep learning architectures. For example, the effectiveness of dense architectures together with different classifiers

was assessed by Mahmoud et al. (2025) using a CT dataset that consisted of 1,000 images related to adenocarcinoma, large-cell carcinoma, squamous-cell carcinoma, and normal classes. In general, their study highlights the promising results of the application of DenseNet-based architectures although the performance of the methods depends on the classifier used. Recently, the study of Yilmaz et al. (2026) considered a more comprehensive comparison of CNN and Vision Transformers using the Chest CT-Scan Images and IQ-OTH/NCCD datasets. The best models obtained accuracy rates of 96.83% and 97.14%, with F1 scores of 0.9692 and 0.9732, respectively.

Recently, other research works have been conducted to examine specific methods for predicting the histological subtypes. In particular, Panchawagh et al. (2025) investigated the role of quantitative computed tomography (CT) radiomics along with machine-learning and deep-learning approaches in the prediction of non-small cell lung cancer (NSCLC) histological subtypes and proved that such non-invasive and quantitative analysis of the CT can be used for the prediction of subtypes. Other modern researches have tried attention-based and multi-scale models to improve feature learning in order to classify lung cancer from the CT image.

Nevertheless, due to the aforementioned developments, the problem of comparing the performance of different CNN architectures is still difficult since the performance that was reported is highly dependent on such factors as the dataset used, the distribution of classes in the dataset, the preprocessing of the data, the training procedure, and the evaluation process itself. The use of different datasets makes it impossible to consider different studies as comparable benchmarks, whereas using the same dataset does not exclude different experimental setups.

In this research paper, the CNN model with the application of transfer learning methodology is being analyzed for classification into four classes including adenocarcinoma, large cell carcinoma, squamous cell carcinoma, and normal lung CT images. The ResNet-50, VGG-16, and Xception models are tested using the same data set but with stratified five-fold cross-validation method. The analysis of performance of the models has been done by means of measures like accuracy, precision, recall, F1-score, and ROC-AUC score. The objective is to identify the most effective architecture among the evaluated CNN models for this dataset and to establish a reproducible comparative baseline for further research in automated lung CT image classification.

Methodology

This study investigates transfer learning-based CNNs for four-class classification of lung computed tomography (CT) images. The target classes are adenocarcinoma, large-cell carcinoma, squamous-cell carcinoma, and normal lung images. Three ImageNet-pretrained CNN architectures, namely ResNet-50, VGG-16, and Xception, were evaluated using a common experimental framework. The overall workflow consists of dataset collection, image preprocessing, stratified five-fold cross-validation, transfer learning-based model construction, model training, and performance evaluation. The overall workflow is illustrated in Figure 1.

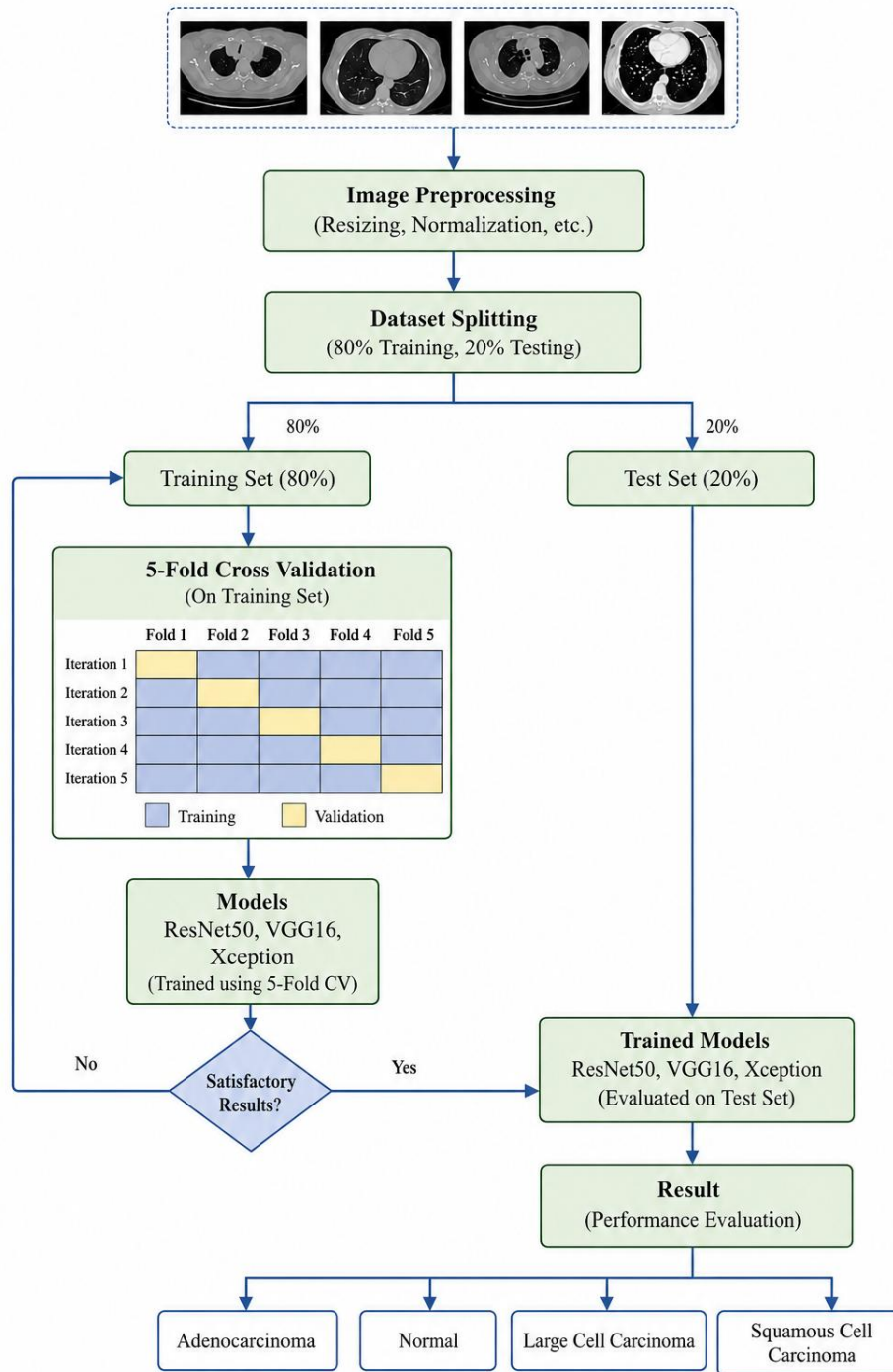


Figure 1. Working flow to classify lung cancer.

Data Collection

Numerous public datasets exist for lung cancer classification tasks, though not all are equally fit for purpose some distinguish only cancerous from non-cancerous cases, while others separate benign from malignant findings without finer subtype detail. Selecting an appropriately labeled dataset is therefore a critical first step toward meaningful results. For this work, CT images were sourced from a publicly hosted Kaggle collection (Hany, 2020), originally compiled from multiple online sources for a chest-cancer detection project combining machine learning and CNN-based methods. Each image's label was manually verified, and rather than being stored in the original DICOM format, the images were converted to JPG or PNG to suit the model pipeline. Overall, the dataset spans three malignant histological categories: adenocarcinoma, large-cell carcinoma, and squamous-cell carcinoma alongside a normal lung category.

Lung adenocarcinoma: This is the most commonly diagnosed lung cancer subtype, making up roughly 40% of NSCLC cases and close to 30% of all lung cancers combined. It typically develops in the peripheral regions of the lung, in tissue associated with air exchange and mucus secretion (Hany, 2020).

Large-cell undifferentiated carcinoma: This subtype has no fixed location within the lung and is marked by aggressive growth and early spread. Representing an estimated 10–15% of NSCLC diagnoses, it tends to progress more rapidly than other subtypes (Hany, 2020).

Squamous cell lung cancer: This subtype tends to originate centrally, near where the trachea meets the larger bronchi or within the major airway branches. It accounts for roughly 30% of NSCLC diagnoses and has a well-established link to smoking history (Hany, 2020).

The full dataset comprised 1,000 images, distributed across classes as follows: 338 adenocarcinoma, 260 squamous-cell carcinoma, 215 normal, and 187 large-cell carcinoma making large-cell carcinoma the least represented class and giving the dataset a moderate class imbalance. Since the source dataset lacked reliable patient-identifier metadata, it was not possible to confirm whether images from the same patient appeared across different splits. As a result, all experiments were carried out at the image level rather than the patient level, and this inability to guarantee patient-independent partitioning is acknowledged as a limitation of the study.

Ethical Considerations: The CT images analyzed in this work came from a publicly hosted chest CT dataset on Kaggle, released without any identifying patient information. Since the research relied entirely on secondary use of already-anonymized public data and involved no direct contact with human subjects, formal ethical clearance and participant consent were deemed unnecessary under our institution's and the journal's guidelines.

Data Preprocessing:

Image preprocessing was performed to ensure that all input images had a consistent spatial size and representation across the evaluated CNN architectures. Each image was resized to 180×180 pixels with three color channels. The 180×180 input size was selected as a computationally efficient resolution that preserves the overall visual patterns and structural characteristics present in the CT images while reducing the computational and memory requirements associated with larger input resolutions. Noteworthy, the same input resolutions were used for ResNet-50, VGG-16, and Xception in order to ensure consistency in comparison between the three different architectures.

The dataset consists of four mutually exclusive categories – adenocarcinoma, large-cell carcinoma, squamous-cell carcinoma, and normal lung. The class labels were one-hot encoded vectors due to the multiclass classification task. There is a four-dimensional binary vector that represents the target labels in the case of four classes, where one represents the corresponding class while all others are zero. Such approach to coding has been used since every image is associated with only one out of the four classes, and thus the one-hot encoding fits perfectly with the Softmax output layer and categorical cross-entropy loss function used in training. The output from the Softmax layer will be the probability distribution across the four mutually exclusive classes, and categorical cross-entropy calculates the difference between such distribution and one-hot encoding.

No augmentations on images, weights on classes, or other types of resampling methods were applied to the experiments. All three algorithms were tested with the same pre-processing and training procedures to ensure a fair comparison between them.

Data Splitting and Five-Fold Cross-Validation

To avoid drawing conclusions from a single, potentially unrepresentative train/validation split, a stratified five-fold cross-validation scheme was adopted. The complete set of 1,000 images was divided into five equally sized folds while preserving the original class proportions in each. During each of the five iterations, four folds (800 images) served as the training set, and the remaining fold (200 images)

was held out for validation — with every fold rotating into the validation role exactly once across the five runs.

Within each fold, the held-out validation images played no role in tuning that fold's model parameters, and the same preprocessing pipeline was applied uniformly across all data regardless of split. No image used for validation in a given fold was ever included in that fold's training set. Performance metrics were computed on each fold's validation set, and the final reported results reflect the mean and standard deviation across all five folds. Because patient identity could not be traced in the source dataset, this cross-validation was necessarily performed at the image level, meaning patient-level independence could not be guaranteed.

Transfer Learning Models:

Three widely used ImageNet-pretrained architectures: VGG-16, ResNet-50, and Xception were selected for evaluation under identical transfer-learning conditions. This selection was deliberate: each network represents a distinct design philosophy, with VGG-16 relying on conventional deep convolutional stacking, ResNet-50 built around residual (skip-connection) learning, and Xception structured around depthwise separable convolutions.

ResNet-50, a 50-layer deep network, introduces shortcut (skip) connections specifically to counter the vanishing-gradient problem that arises in very deep architectures. These connections enable smoother feature propagation through the network and have contributed to strong results across image classification, object detection, and medical imaging applications (He et al., 2016; Sharma & Guleria, 2022).

VGG-16 is a 16-layer network built primarily from repeated 3×3 convolution and max-pooling blocks. Its straightforward, uniform structure supports the extraction of hierarchical features and has made it a common baseline choice in medical imaging tasks, largely due to its consistent and reliable feature-learning behavior (Simonyan & Zisserman, 2015; Ajit et al., 2020).

Xception builds on the Inception architecture family by substituting standard convolutional layers with depth wise separable convolutions, which decouple spatial and channel-wise feature learning. This substitution lowers computational demands considerably without sacrificing representational power, making Xception a computationally efficient option for image classification (Chollet, 2017).

All three backbones were initialized using pretrained ImageNet weights, which remained frozen throughout training. Each model's original ImageNet classification head was swapped out for an identical custom fully connected head, so that any performance gaps observed between models could be attributed to differences in backbone feature extraction rather than variation in classifier architecture.

Model Architecture and Training:

A controlled comparison was maintained by training all three architectures with an identical classification-head design and training setup, after removing each backbone's original ImageNet classification layer. In ResNet-50's case, global average pooling within the backbone yielded a 2,048-dimensional feature vector that was then passed into the custom classification layers, with the underlying pretrained feature extractor left frozen throughout training.

This feature vector then flowed through three fully connected layers of decreasing size — 3,000, 1,500, and 750 neurons — each using ReLU activation to inject nonlinearity into the learned representations. A final four-neuron dense layer, corresponding to the four target classes, applied a Softmax function to output class probabilities. The complete ResNet-50 architecture with the custom classification head is illustrated in Figure 2.

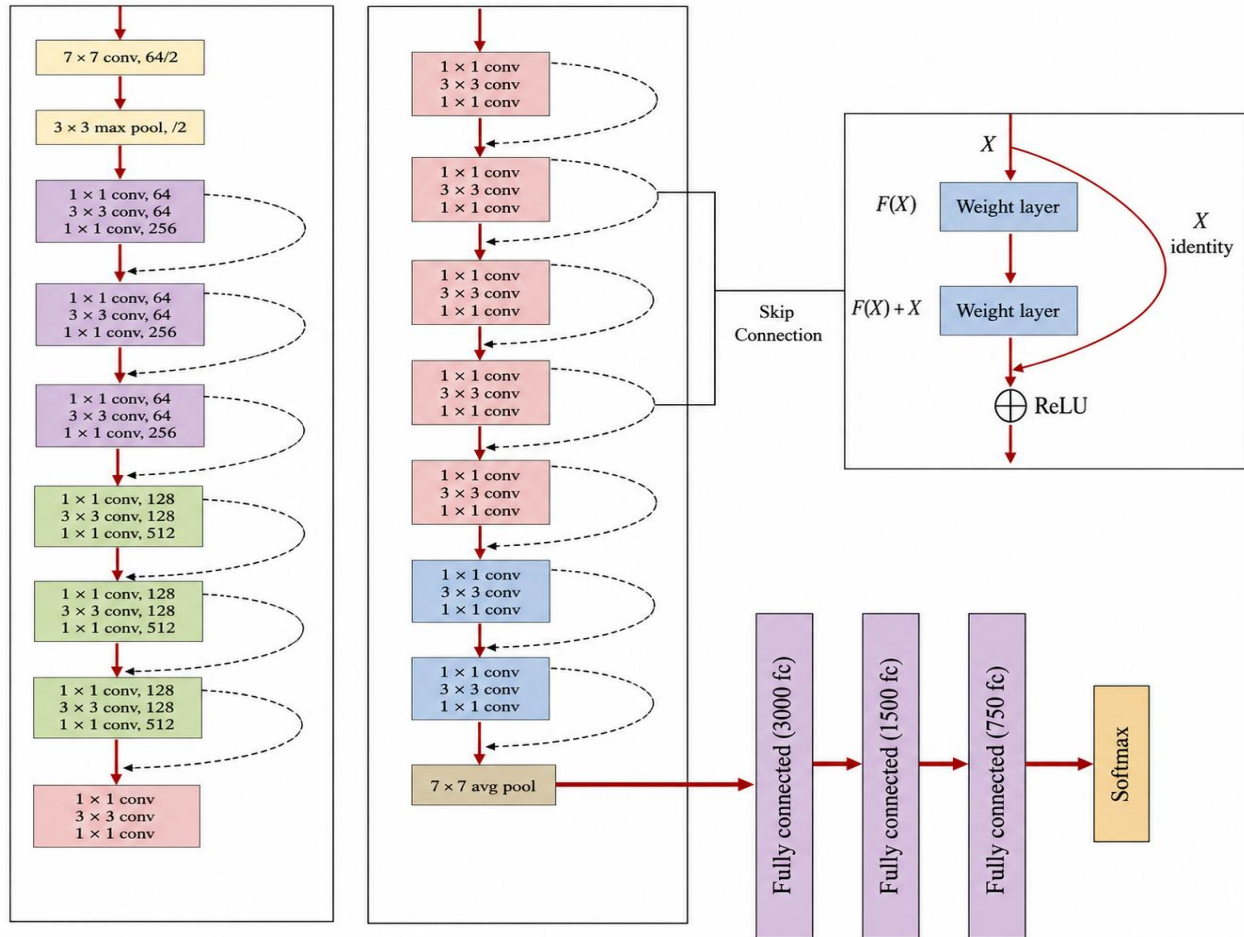


Figure 2. Resnet-50 model architecture with custom classification head

Applying this same classification-head configuration across ResNet-50, VGG-16 and Xception minimized architectural variability at the classifier stage. Training ran for 25 epochs with a batch size of 32, using the Adam optimizer at a 0.001 learning rate and categorical cross-entropy as the loss function. Backbone fine-tuning was deliberately excluded, meaning training updates applied only to the newly introduced classification layers. Table I summarizes the common training configuration used for all three models.

Table 1. Hyper parameters used for training the models

Parameter Name	Assigned Value
Batch Size	32
Epochs	25
Learning Rate	0.001
Input Image Size	180×180×3
Loss	Categorical Cross-Entropy
Optimizer	Adam
Hidden Layer Activation	ReLU
Activation Function	Softmax
Number of Output Classes	4
Pretrained Weights	ImageNet
Cross-validation	Stratified 5-Fold

Model Evaluation

The performance of each individual classifier was independently evaluated based on the held out validation fold within each cross-validation iteration through five evaluation criteria: accuracy, macro-averaged precision, macro-averaged recall, macro-averaged F1 score, and ROC-AUC. Accuracy is defined as the total number of images that are classified correctly. Precision is the ratio of correctly predicted positives to all positive predictions, whereas recall is the ratio of all positively predicted items to the actual number of positive items. F1 score is the harmonic average of precision and recall. All three metrics (precision, recall, and F1 score) were macro-averaged such that all classes have the same importance.

Calculation of the ROC-AUC was done using the one-vs-rest approach for determining the separability of the classes from the rest. Macro- and weighted-averaging versions are both reported, whereby macro averaging gives equal importance to all the classes while the weighted version accounts for different numbers of samples in different classes. Performance is measured in terms of the mean and the standard deviation over all five folds. Predictions for all the five folds of validation are combined to give the out-of-fold (OOF) dataset. This provides aggregate metrics as well as confusion matrix. This is a complete picture of the classification since each image only gets one OOF prediction, which must be generated using the fold that did not have the image during training.

Results and Discussion

The findings on the evaluations done for ResNet-50, VGG-16, and Xception when used to classify lungs CT scans are provided in this part of the study. To have a more reliable measure of the performance of the models, stratified five-fold cross-validation was used. In each fold, there were 800 images used for training and 200 images for validation. The same configuration for training was used for all three models. The results are reported as mean $\pm$ standard deviation across the five folds.

Five-Fold Cross-Validation Results:

Table II presents the fold-wise performance of the ResNet-50 model. The model achieved validation accuracies of 95.00%, 94.50%, 93.50%, 91.50%, and 89.00% across the five folds, respectively. The corresponding mean accuracy was 92.70% with a standard deviation of 2.46 percentage points.

Table II. Fold-wise Performance of ResNet-50

Fold	Accuracy (%)	Macro Precision (%)	Macro Recall (%)	Macro F1-score (%)	Macro ROC-AUC (%)	Weighted ROC-AUC (%)
1	95.00	95.72	94.25	94.84	98.40	98.58
2	94.50	95.37	94.05	94.58	99.19	99.18
3	93.50	94.40	93.42	93.86	99.16	99.09
4	91.50	93.46	91.18	92.12	99.05	98.94
5	89.00	89.48	89.51	89.37	98.29	98.11
Mean $\pm$ SD	92.70 $\pm$ 2.46	93.69 $\pm$ 2.51	92.48 $\pm$ 2.06	92.95 $\pm$ 2.27	98.82 $\pm$ 0.44	98.78 $\pm$ 0.44

The results show that ResNet-50 maintained an accuracy above 89% in all five validation folds. The highest accuracy of 95.00% was obtained in Fold 1, whereas the lowest accuracy of 89.00% was obtained in Fold 5. The standard deviation of 2.46 percentage points indicates moderate variation in performance across the validation folds. Reporting the mean and standard deviation therefore provides a more representative assessment than reporting only the highest-performing fold. The macro-averaged

metrics were also consistently high across the five folds. ResNet-50 achieved a mean macro precision of $93.69 \pm 2.51\%$, macro recall of $92.48 \pm 2.06\%$, and macro F1-score of $92.95 \pm 2.27\%$. The mean macro ROC-AUC was $98.82 \pm 0.44\%$, while the corresponding weighted ROC-AUC was $98.78 \pm 0.44\%$.

Comparison of the Evaluated CNN Architectures

The three evaluated architectures were trained using the same dataset, preprocessing procedure, classification-head structure, optimizer, learning rate, batch size, number of epochs, and frozen-backbone strategy. This common experimental configuration allows their classification performance to be compared under the same conditions. Table III presents the overall performance comparison of the evaluated models.

Table III. Overall Performance Comparison of the Evaluated Models

Model	Accuracy (%)	Macro Precision (%)	Macro Recall (%)	Macro F1-score (%)	Macro ROC-AUC (%)
ResNet-50	92.70 ± 2.46	93.69 ± 2.51	92.48 ± 2.06	92.95 ± 2.27	98.82 ± 0.44
Xception	89.00 ± 2.15	89.63 ± 1.88	89.30 ± 2.22	89.32 ± 2.05	97.71 ± 1.11
VGG-16	83.70 ± 2.97	87.66 ± 1.67	84.50 ± 3.20	84.02 ± 3.00	97.91 ± 0.44

ResNet-50 achieved the highest performance across the primary classification metrics. Its mean accuracy of $92.70 \pm 2.46\%$ was 3.70 percentage points higher than Xception and 9.00 percentage points higher than VGG-16. Similarly, ResNet-50 achieved the highest macro precision, macro recall, and macro F1-score among the three architectures. The results indicate that the residual feature representation learned by the ResNet-50 backbone was more effective for the present four-class CT classification task under the selected experimental configuration. Xception achieved intermediate performance, while VGG-16 produced the lowest mean accuracy and macro F1-score. Although VGG-16 obtained a macro ROC-AUC of 97.91%, slightly higher than the 97.71% obtained by Xception, its classification accuracy and macro F1-score were considerably lower. This demonstrates that ROC-AUC and accuracy measure different aspects of model performance. ROC-AUC evaluates discrimination across classification thresholds, whereas accuracy evaluates the final class assignments.

Out-of-Fold Class-wise Performance of ResNet-50

To further analyze the classification behavior of ResNet-50, predictions from all five validation folds were combined to generate out-of-fold (OOF) predictions for the complete dataset. Each image was predicted by a model that had not used that image for training within the corresponding fold. The pooled OOF results produced an accuracy of 92.70%, macro precision of 93.50%, macro recall of 92.49%, and macro F1-score of 92.94% as demonstrated in Table IV. The OOF macro ROC-AUC and weighted ROC-AUC were 98.84% and 98.80%, respectively.

Table IV. Class-wise Out-of-Fold Performance of ResNet-50

Class Name	Precision (%)	Recall (%)	F1 Score (%)	Support
Adenocarcinoma	88.06	93.79	90.83	338
Large Cell Carcinoma	93.14	87.17	90.06	187
Normal	99.53	98.60	99.07	215
Squamous Cell Carcinoma	93.25	90.38	91.80	260
Macro average	93.50	92.49	92.94	1000

The class-wise results indicate that the model performed particularly well on the normal category, which achieved a precision of 99.53%, recall of 98.60%, and F1-score of 99.07%. Among the malignant subtypes, adenocarcinoma achieved the highest recall at 93.79%, while squamous-cell carcinoma achieved the highest F1-score at 91.80%. Large-cell carcinoma produced the lowest recall among the four classes at 87.17%, indicating that this subtype was comparatively more difficult for the model to identify. Nevertheless, its precision was 93.14%, suggesting that predictions assigned to the large-cell

carcinoma class were generally reliable. Overall, the class-wise results show that the main classification challenge was distinguishing among malignant subtypes rather than distinguishing malignant images from normal lung CT images.

Confusion Matrix Analysis

The aggregated OOF confusion matrix for ResNet-50 is presented in Table V. The confusion matrix provides further insight into the model's classification behavior. Of the 338 adenocarcinoma images, 317 were correctly classified, while 9 were classified as large-cell carcinoma, 1 as normal, and 11 as squamous-cell carcinoma. For large-cell carcinoma, 163 of 187 images were correctly classified, while 18 were classified as adenocarcinoma and 6 as squamous-cell carcinoma. The normal class was classified with high consistency, with 212 of 215 images correctly identified.

Table V. Out-of-Fold Confusion Matrix of ResNet-50

Actual / Predicted	Adenocarcinoma	Large-cell carcinoma	Normal	Squamous-cell carcinoma
Adenocarcinoma	317	9	1	11
Large-cell carcinoma	18	163	0	6
Normal	3	0	212	0
Squamous-cell carcinoma	22	3	0	235

Only three normal images were classified as adenocarcinoma, and none were classified as large-cell or squamous-cell carcinoma. For squamous-cell carcinoma, 235 of 260 images were correctly classified. Twenty-two were classified as adenocarcinoma and three as large-cell carcinoma. The largest source of misclassification was therefore the confusion between squamous-cell carcinoma and adenocarcinoma. These results suggest that the model learned to have capabilities of differentiating between healthy and malignant cases, although differentiation within malignant cases proved more difficult.

Training and Validation Behavior

The patterns for training and validation accuracy offer insights into how ResNet-50 learned over the course of each of the five cross-validation folds. As shown in Figure 3, the training accuracy rose quickly and came close to reaching near-100% by the end, but validation accuracy remained lower and demonstrated variability across folds. It is clear from the graph that there was a visible difference in training and validation accuracies in some of the folds, indicating that there was some overfitting to the training data for the newly trained classification head. This is especially true for Fold 5, where training accuracy reached close to 100%, but the final validation accuracy was only about 89%. However, the validation accuracy stayed above 89% in all five folds. The fold-wise graphs also show that the learning pattern was not consistent for the validation partitions, just like the differences in the validation accuracy reported in Table II, which ranged from 89.00% to 95.00%.

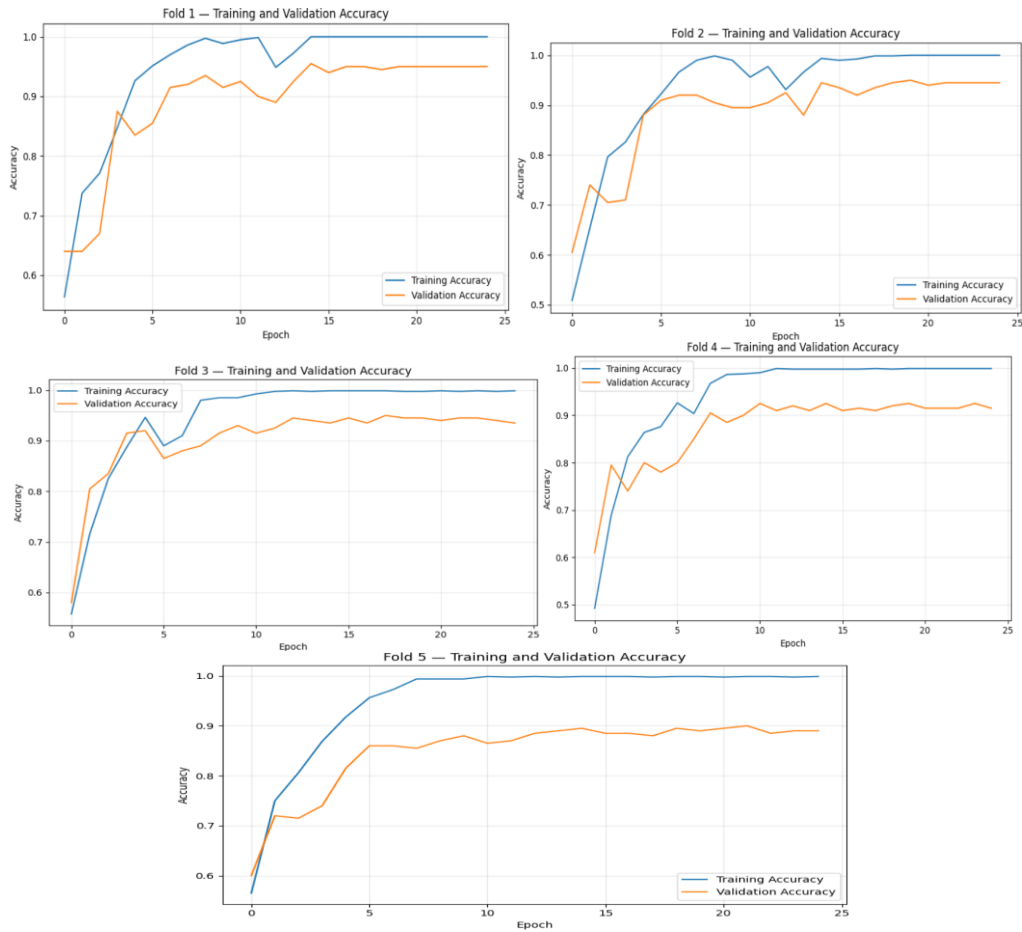
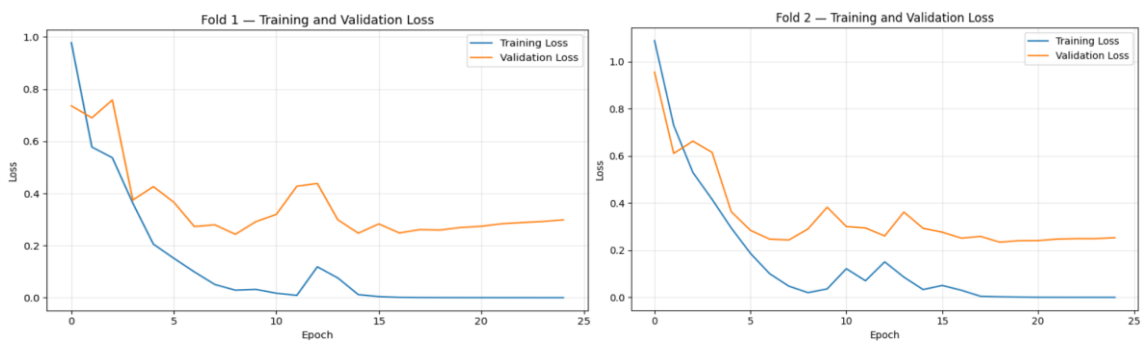


Figure 3. Training and validation accuracy curves of ResNet-50 across the five stratified cross-validation folds

Training and validation loss curves provide information about the learning process of ResNet-50 in five cross-validation runs. The behavior of the training loss curve can be seen in Figure 4 where a continuous drop of the loss to zero was observed as the training progressed. In comparison, validation loss first dropped as the number of epochs increased but then stabilized or slightly increased in some folds. Such behavior of the training and validation loss is associated with the slight overfitting phenomenon shown by the accuracy curves where the classifier continued to learn the training set even though validation performance started to flatten out. Despite the variations, validation loss remained stable in all five folds and was supported by a high validation accuracy.



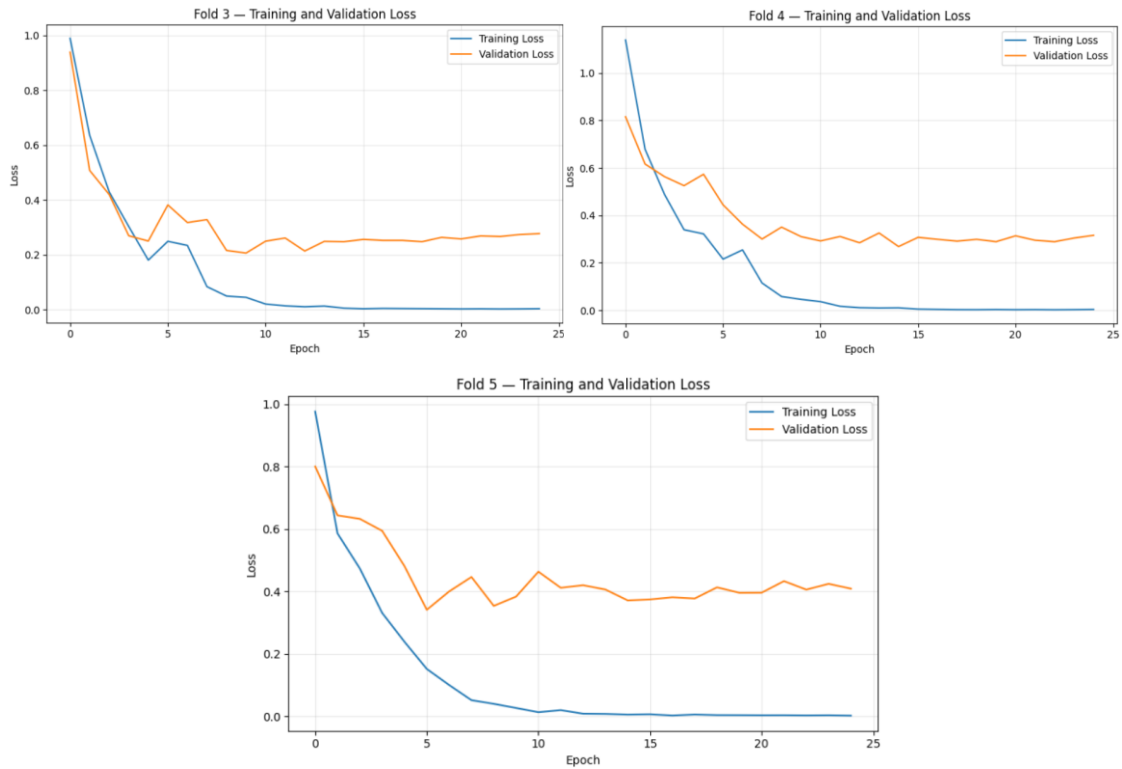
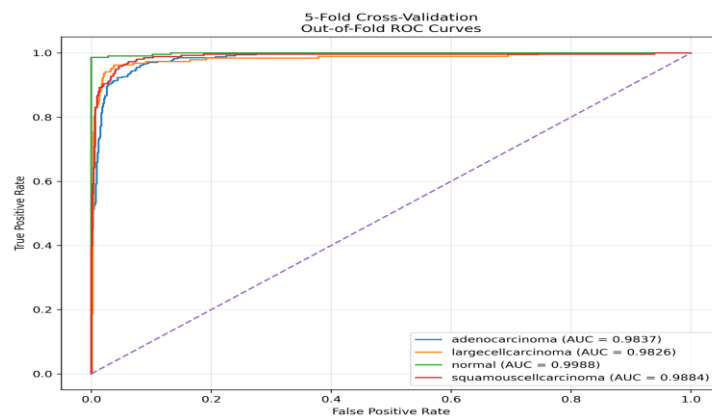


Figure 4. Training and validation loss curves of ResNet-50 across the five stratified cross-validation folds

ROC-AUC Analysis:

The ROC analysis was done in order to check how discriminative the examined convolutional neural networks (CNNs) architectures were. In Figure 5 one can see ROC curves generated using out-of-fold predictions of ResNet-50, VGG-16, and Xception on the whole dataset. ResNet-50 achieved the highest mean macro ROC-AUC of $98.82 \pm 0.44\%$, followed by VGG-16 at $97.91 \pm 0.44\%$ and Xception at $97.71 \pm 1.11\%$. The high values of ROC-AUC prove good discrimination power of all considered architectures in accordance with the adopted cross-validation protocol, with the best results demonstrated by ResNet-50. When interpreting the ROC-AUC values, one should consider them together with the accuracy and F1-score metrics described in Table III.

In particular, relatively high ROC-AUC values for VGG-16 and Xception in spite of their poor classification accuracy show that ROC-AUC and accuracy are quite different indicators. While ROC-AUC evaluates discrimination on different thresholds, accuracy is the indicator of the final class's estimation.



(a) ResNet-50

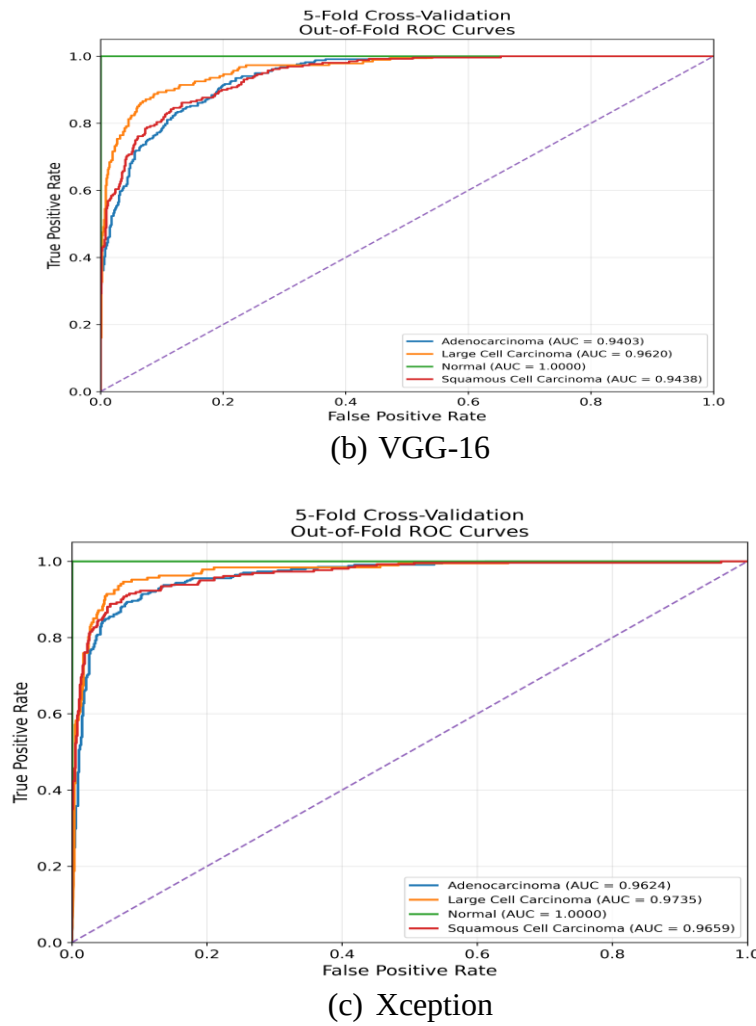


Figure 5. AUC-ROC curves of the models: (a) ResNet-50, (b) VGG-16, (c) Xception

Comparison with Previous Studies

Table VI places the method under review in context with other relevant research works on image classification for lung cancer. Since performance depends on both the data and evaluation methodology, the works discussed below are classified according to data similarity.

From among the studies cited in Table VI, Muntasir et al. (2023) and Mahmoud et al. (2025) use the same Chest CT-Scan Images data from Kaggle. Muntasir et al. (2023) made use of the InceptionV3, ResNet-50, and Xception models and obtained accuracy scores of 92%, recall of 91.72%, and AUC of 98.21%. On the contrary, Mahmoud et al. (2025) evaluated various DenseNet models with classifiers and obtained an accuracy score of 83%. In comparison, the evaluated ResNet-50 achieved $92.70 \pm 2.46\%$ accuracy, $92.95 \pm 2.27\%$ macro F1-score, and $98.82 \pm 0.44\%$ macro ROC-AUC using stratified five-fold cross-validation.

The algorithm being analyzed is able to perform competitively using the same dataset. It further shows the average and the standard deviation of the results from five folds, thus providing a more reliable measure of stability than just one train/test split. However, direct numerical comparison should still be made carefully because the previous studies may have used different data splits, preprocessing methods, augmentation techniques, and training settings.

The remaining studies in Table VI used different datasets. Tekade et al. (2018) reported 91.66% accuracy using CNN on LIDC-IDRI, while Ausawalaithong et al. (2018) achieved $74.43 \pm 6.01\%$ accuracy using CNN on chest X-ray images. Ali et al. (2018) reported 64.4% accuracy on LIDC-IDRI using reinforcement learning. Pandian et al. (2022) achieved 83% accuracy and an F1-score of 0.89 on CT-scan images. More recently, Yilmaz et al. (2026) reported accuracies of 96.83% and 97.14% on the

Chest CT-Scan Images and IQ-OTH/NCCD datasets, respectively, using CNN- and Vision Transformer-based models.

Table VI. Comparison with Previous Lung CT Classification Studies

Authors (Year)	Applied Models	Dataset Collection	Performance Metrics
Tekade et al. (2018)	SVM, CNN	LIDC-IDRI	SVM - Accuracy: 90%, CNN - Accuracy: 91.66%
Ausawalaithong et al. (2018)	CNN	JSRT, Chest X-ray	Accuracy: $74.43 \pm 6.01\%$, Specificity: $74.96 \pm 9.85\%$, Sensitivity: $74.68 \pm 15.33\%$
Ali et al. (2018)	Reinforcement Learning	LIDC-IDRI	Accuracy: 64.4%, Sensitivity: 58.9%, Specificity: 55.3%, PPV: 54.2%, NPV: 60.0%
Pandian et al. (2022)	CNN, Google Net	CT-scan images	Accuracy: 83%, F1 Score: 0.89
Muntasir et al. (2023)	CNN, Inception V3, ResNet50, Xception	The Chest CT-Scan Images (Kaggle)	Accuracy: 92%, Recall: 91.72%, AUC: 98.21%, Loss: 0.328
Mahmoud et al. (2025)	DenseNet variants with Classifiers	The Chest CT-Scan Images (Kaggle)	Train Accuracy: 92% Test Accuracy: 83%
Yilmaz et al. (2026)	CNN-based models, ViT-based models	The Chest CT-Scan Images (Kaggle), IQ-OTH/NCCD	Accuracy: 96.83% & 97.14%, F1-score: 0.9692, 0.9732
Our work (2026)	ResNet-50, VGG-16, Xception	The Chest CT-Scan Images (Kaggle)	ResNet-50 Accuracy: $92.70 \pm 2.46\%$; Macro F1-score: $92.95 \pm 2.27\%$; Macro ROC-AUC: $98.82 \pm 0.44\%$

A comparison with the most recent study by Yilmaz et al. (2026) further clarifies the difference between the two approaches. Yilmaz et al. evaluated a broad range of CNN and Vision Transformer architectures, whereas the present study focuses on three established transfer-learning CNNs: ResNet-50, VGG-16, and Xception. Their study used both the Chest CT-Scan Images and IQ-OTH/NCCD datasets, while the present study uses the Chest CT-Scan Images dataset. Yilmaz et al. reported an F1-score of 0.9732 for their best-performing model, whereas the evaluated ResNet-50 achieved a macro F1-score of $92.95 \pm 2.27\%$. However, these results should not be directly ranked because the studies used different experimental settings. In particular, the present study uses stratified five-fold cross-validation and reports mean $\pm$ standard deviation across folds. Thus, the main distinction of the present work is its controlled comparison of established CNN architectures under a common evaluation protocol, rather than the introduction of a new network architecture.

These results provide useful context but are not direct benchmarks because the datasets and experimental settings differ. Thus, the relevant comparison will be with those studies that use the same Kaggle dataset. In this regard, the ResNet-50 model under analysis shows decent results by having high ROC-AUC and low variance between folds. It also provides the best performance among all three models under investigation.

Discussion

Of the three transfer learning architectures examined, ResNet-50 shows the best results for the present classification problem: $92.70 \pm 2.46\%$ mean accuracy, $92.95 \pm 2.27\%$ macro F1-score, and $98.82 \pm 0.44\%$ macro ROC-AUC. The results show good performance under the examined conditions. The reason for the better performance of ResNet-50 compared to VGG-16 and Xception can be associated with the residual learning framework of ResNet-50 and its ability to pass features using shortcut connections. Since all three models were examined using the same architecture and the same setting, it provides empirical evidence for comparison of the pretrained feature representation of the present dataset. According to the confusion matrix, most of the misclassifications occur in malignant subtypes, namely, adenocarcinoma and squamous-cell carcinoma, but not in the normal class, which is recognized much more accurately. It means that the model recognizes well healthy tissue vs malignant tissue, but not so well different subtypes of malignancy. In terms of five-fold cross-validation, ResNet-50 shows an accuracy from 89.00% to 95.00% (SD = 2.46%). This spread underscores why reporting mean $\pm$ standard deviation, rather than a single figure, gives a fairer picture of model performance. That said, a few limitations temper how these results should be read. The dataset consists of publicly available JPG/PNG images collected from multiple sources and does not provide sufficient patient-level identifiers to verify patient-independent partitioning. Therefore, the present evaluation represents image-level stratified cross-validation rather than confirmed patient-level cross-validation. Additionally, validation relied entirely on the same dataset, with no independent external clinical data involved. The performance figures reported here therefore reflect computational classification accuracy on this specific dataset rather than clinical diagnostic accuracy, and should not be taken as evidence that the model could substitute for a clinical or pathological diagnosis. External validation using larger datasets with reliable patient-level identifiers and diverse imaging characteristics is required to assess the generalizability of the approach. Overall, the results demonstrate that transfer learning can provide effective feature representations for multiclass lung CT image classification. ResNet-50 produced the best overall performance among the three evaluated architectures under the common experimental configuration, while the class-wise and confusion-matrix analyses highlight the remaining challenge of distinguishing between malignant lung cancer subtypes.

Conclusion

This work put forward a deep-learning pipeline for classifying lung CT images into four categories, centered on ResNet-50 and benchmarked against VGG-16 and Xception. ResNet-50 came out ahead, delivering a mean accuracy of $92.70 \pm 2.46\%$ and a macro F1-score of $92.95 \pm 2.27\%$. The above-mentioned results suggest that transfer learning can prove useful for the classification of CT images based on lung cancer subtypes in the current dataset. There are several drawbacks worth mentioning. First, the dataset was rather small and consisted of data available on the public platform where patient identification is not possible to ensure patient-level independence. Second, there were no other publicly available datasets featuring four classes that would allow for external validation. This means that the current evaluation was performed through stratified cross-validation. In order to conduct future research, it would be desirable to perform validation on larger datasets containing different centers and ensuring patient-level independence as well as labeling. Moreover, domain adaptation and label harmonization can be considered to test the generalizability across datasets of different classes. Further enhancements can include data augmentation, fine-tuning, applying attention and explainable AI.

Conflicts of Interest: The authors declare no competing interests.

References

- Ajit, A., Acharya, K., and Samanta, A., 2020. A review of convolutional neural networks. In *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, pp. 1–5.
- Ali, I., Hart, G. R., Gunabushanam, G., Liang, Y., Muhammad, W., Nartowt, B., Kane, M., Ma, X., and Deng, J., 2018. Lung nodule detection via deep reinforcement learning. *Frontiers in Oncology*, 8.
- Ausawalaithong, W., Thirach, A., Marukatat, S., and Wilaiprasitporn, T., 2018. Automatic lung cancer prediction from chest X-ray images using the deep learning approach. In *2018 11th Biomedical Engineering International Conference (BMEiCON)*, IEEE.
- Chhikara, B. S.; Parang, K. Global Cancer Statistics 2022: The Trends Projection Analysis. *Chem Biol Lett* 2023, 10 (1), 451.
- Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, IEEE Computer Society, pp. 1800–1807.
- Faisal, M., Bashir, S., Khan, Z., and Khan, F., 2018. An evaluation of machine learning classifiers and ensembles for early stage prediction of lung cancer.
- Hany, M., 2020. *Chest CT-Scan images Dataset*. Kaggle. [Online]. Available: <https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images> (accessed 21 September 2023).
- He, K., Zhang, X., Ren, S., and Sun, J., 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.
- Herbst, R. S., Heymach, J. V., and Lippman, S. M., 2008. Lung cancer. *New England Journal of Medicine*, 359(13), pp. 1367–1380.
- Kaur, S., Hooda, R., Mittal, A., Akashdeep, and Sofat, S., 2017. Deep CNN-based method for segmenting lung fields in digital chest radiographs. In *Advanced Informatics for Computing Research: First International Conference, ICAICR 2017*, Springer, pp. 185–194.
- Mahmoud, M. A., Wu, S., Su, R., Wen, Y., Liu, S., and Guan, Y., 2025. Classifiers combined with DenseNet models for lung cancer computed tomography image classification: A comparative analysis. *Current Medical Imaging*, 21.
- Muntasir, M., Mahmud, M. I., Meherin, M., and Abdelgawad, A., 2023. LCD-CTCNN: Lung cancer diagnosis of CT scan images using CNN based model. In *2023 10th International Conference on Signal Processing and Integrated Networks (SPIN)*, IEEE, pp. 205–212.
- Nadkarni, N. S., and Borkar, S., 2019. Detection of lung cancer in CT images using image processing. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 863–866.
- Nazir, I., AlQahtani, S. A., Jadoon, M. M., Dahshan, M., et al., 2023. Machine learning-based lung cancer detection using multiview image registration and fusion. *Journal of Sensors*, 2023.
- Panchawagh, S., Halder, A., Haldule, S., Sanker, V., Lalwani, D., Sequeria, R., Naik, H., and Desai, A., 2025. Quantitative radiomic analysis of computed tomography scans using machine and deep learning techniques accurately predicts histological subtypes of non-small cell lung cancer: A retrospective analysis. *European Journal of Surgical Oncology*, 51(10), p. 110376.
- Pandian, R., Vedanarayanan, V., Ravi Kumar, D., and Rajakumar, R., 2022. Detection and classification of lung cancer using CNN and Google Net. *Measurement: Sensors*, 24, p. 100588.
- Radhika, P. R., Nair, R. A., and G., V., 2019. A comparative study of lung cancer detection using machine learning algorithms. In *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pp. 1–4.
- Sharma, S., and Guleria, K., 2022. Deep learning models for image classification: comparison and applications. In *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, IEEE, pp. 1733–1738.
- Simonyan, K., and Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*

- Sung, H., Ferlay, J., Siegel, R., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F., 2021. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(2).
- Tekade, R., 2018. Lung nodule detection and classification using machine learning techniques. *Asian Journal for Convergence in Technology (AJCT)*, 4(April).
- Yilmaz, M. T., Algul, E., and Pacal, I., 2026. Robust artificial intelligence frameworks for lung cancer subtyping and malignancy detection on thoracic CT. *Scientific Reports*, 16, p. 27223.