Research Article GU J Sci, Part A, 12(2): 373-391 (2025) 10.54287/gujsa.1648772

Gazi University
Journal of Science
PART A: ENGINEERING AND INNOVATION
http://dergipark.org.tr/gujsa

Deep Learning-Based Lung Cancer Diagnosis: Data Balancing, Model Optimisation and Performance Analysis

Feyyaz ALPSALAZ[1*]

[1] *Yozgat Bozok University, Yozgat, Türkiye*

Keywords

Lung Cancer
Deep Learning
Computed Tomography
CNN
ResNet101

Abstract

Lung cancer (LC) is one of the most lethal malignancies worldwide, and early detection is essential. This study develops a deep learning (DL) based classification model for LC diagnosis using computed tomography (CT) images. In the experiments conducted on the IQ-OTHNCCD LC dataset, the Synthetic Minority Over-sampling Technique (SMOTE) method was applied to eliminate class imbalance, data augmentation techniques were used, and an early stopping mechanism was integrated to enhance the model's generalizability. Commonly used convolutional neural network (CNN) architectures, such as ResNet101, VGG19, and DenseNet121, are compared, and the model's performance is analyzed in detail. With an accuracy of 98%, the trial results demonstrate that the suggested ResNet101 model offers the best classification performance. the DenseNet121 model exhibited a relatively lower accuracy rate in distinguishing between benign and normal classes. The study conclusively demonstrates that an optimized ResNet101-based deep learning model, enhanced with data balancing and augmentation techniques, provides the most accurate and reliable classification performance for lung cancer detection using CT images. It not only outperforms traditional CNN architectures in terms of overall accuracy (98%) but also achieves perfect classification in malignant cases. These results validate the model's potential as a robust diagnostic aid and highlight its superiority over existing methods in the literature, particularly in handling class imbalance and maintaining generalization across diverse image types.

Cite

Alpsalaz, F. (2025). Deep Learning-Based Lung Cancer Diagnosis: Data Balancing, Model Optimisation and Performance Analysis. *GU J Sci, Part A*, *12*(2), 373-391. doi:10.54287/gujsa.1648772

Author ID (ORCID Number)

0000-0002-7695-6426 Feyyaz ALPSALAZ

Article Process

Submission Date 28.02.2025
Revision Date 21.03.2025
Accepted Date 25.03.2025
Published Date 30.06.2025

1. INTRODUCTION

Lung cancer (LC) is one of the world's most aggressive and deadly malignancies, responsible for approximately 1.8 million deaths annually according to recent WHO data (Thandra et al., 2021). Its high mortality rate is largely due to the lack of symptoms in the early stages, which often delays diagnosis and treatment. Early detection therefore plays a critical role in improving patient survival.

Computed tomography (CT) has become the most widely used modality in the diagnosis and follow-up of LC due to its ability to provide high-resolution cross-sectional images of lung tissue. CT scans allow clinicians to detect lung nodules, differentiate between benign and malignant lesions, and assess disease progression (Tárnoki et al., 2024). However, conventional diagnostic processes rely heavily on manual interpretation by radiologists, which can be time-consuming and prone to variability due to human subjectivity and fatigue.

*Corresponding Author, e-mail: feyyaz.alpsalaz@bozok.edu.tr

In recent years, the field of medical image analysis has witnessed transformative advances with the introduction of DL techniques. DL-based approaches have enabled the development of automated diagnostic systems that significantly improve accuracy while reducing time and error rates in clinical workflows (Özdemir et al., 2025). These models, particularly convolutional neural networks (CNNs), are capable of learning complex spatial features from CT images - from low-level textures to high-level structural patterns - that aid accurate diagnosis (Wang, 2022). When trained on large datasets, DL models outperform traditional methods by providing consistent, fast and objective assessments that are often superior to human interpretation (Javed et al., 2024).

Despite these advances, many current DL-based studies focus narrowly on classification tasks, overlooking challenges such as data imbalance and overfitting. The novelty of this study lies in its integrated approach, which includes preprocessing, class balancing using the Synthetic Minority Over-sampling Technique (SMOTE), and model optimisation strategies to improve performance across classes. The study uses a dataset of 1,190 CT scan slices from the IQ-OTH/NCCD dataset, categorised into benign, malignant and normal classes, and applies three powerful CNN architectures: ResNet101, VGG19 and DenseNet121.

The rest of this paper is structured as follows: Section 2 presents a review of related studies in the literature. Section 3 describes the dataset characteristics, the preprocessing pipeline, the model configurations, and the evaluation metrics. Section 4 discusses the experimental results, including model comparisons and performance analysis. Finally, Section 5 summarises the main findings and suggests future research directions.

2. LITERATURE REVIEW

In recent years, there has been an increasing amount of research focusing on the application of DL methods for LC diagnosis, particularly using CT images. CNNs, with their high accuracy and automatic feature extraction capabilities, have become the dominant architecture in these studies.

A number of studies have proposed hybrid or explainable DL models. Wani et al. (2024) developed DeepXplainer, a DL model that incorporates explainable artificial intelligence (XAI) techniques for LC diagnosis. This model optimises cancer diagnosis with the combination of CNN and XGBoost, and makes the model outputs explainable with the SHAP method. In the study, 97.43% accuracy, 98.71% sensitivity and 98.08% F1 score were obtained (Wani et al., 2024). Similarly, Mohamed et al. (2023) propose a hybrid CNN model with Ebola Optimisation Search Algorithm (EOSA) for LC diagnosis. The model was tested on the IQ-OTH/NCCD dataset and the EOSA CNN achieved 93.21% accuracy, 90.38% sensitivity and 97.95% specificity. The hyper-parameter optimisation of the EOSA algorithm improved the classification performance of the CNN model (Mohamed et al., 2023).

Several studies have combined imaging modalities. Rajasekar et al. (2023) developed a DL based system for LC diagnosis using CT and histopathological images. The results show that CNN-based approaches provide

high accuracy in distinguishing malignant from benign lesions in CT images (Rajasekar et al., 2023). Mamatha et al. (2023) developed a DL model for LC detection using CT and histopathological images. In the study, the diagnostic performance was improved by using trained CNN models including VGG-19 and ResNet-50. The model showed superior performance compared to traditional approaches in terms of sensitivity, specificity and F1 score (Mamatha et al., 2023). Devarajan et al. (2023) developed a DL model for automatic LC detection using CT and chest X-ray images. In the study, the use of data augmentation and transfer learning techniques significantly improved model performance and demonstrated the superiority of CNN-based approaches, particularly in early detection (Devarajan et al., 2023). Davri et al. (2023) systematically reviewed DL studies using histological and cytological images for LC diagnosis, prognosis and prediction. The study analysed the effectiveness of different CNN models in differentiating between adenocarcinoma, squamous cell carcinoma and small cell LC, and highlighted that DL-based models improve diagnostic accuracy (Davri et al., 2023). Zhang et al. (2024) developed a DL model using histopathological images to predict prognosis and treatment response in patients with small cell LC (SCLC). The model predicted patient survival by identifying 50 histomorphological phenotype clusters using contrastive clustering. These results demonstrate that DL analysis of pathology images can make an important contribution to clinical decision making (Zhang et al., 2024).

Multimodal and multi-omics approaches have also gained traction. Sangeetha et al. (2024) propose an improved model for LC diagnosis using a deep neural network based on multimodal data fusion (MFDNN). By combining genetic, clinical and image data, the MFDNN model achieved 92.5% accuracy, 87.4% precision and 86.4% sensitivity, higher than previous models. The study shows that multimodal data integration can make a significant contribution to LC diagnosis (Sangeetha et al., 2024). Tran et al. (2024) developed a DL-based decision support system using genomic and proteomic data in LC diagnosis and treatment. The study contributes to LC biomarker discovery and personalised treatment planning by enabling the identification of cancer subtypes through multi-omics data fusion (Tran et al., 2024).

Advanced CNN architectures and 3D modelling have also been explored. Crasta et al. (2023) propose a new DL model based on 3D-VNet and 3D-ResNet for LC diagnosis. The model achieved successful results in segmentation and classification with 99.34%. These results demonstrate the importance of 3D CNN models in LC detection (Crasta et al., 2024). Said et al. (2023) developed a UNETR-based model for image segmentation in LC diagnosis. This model achieved 97.83% segmentation accuracy and 98.77% classification accuracy by using self-supervised learning techniques to discriminate between malignant and benign nodules in 3D CT images (Said et al., 2023).

Risk prediction and survival analysis have been studied by Mikhael et al. (2023) developed a DL model called Sybil that predicts future LC risk from a single low-dose CT (LDCT) scan. The model predicted cancer development within 1 year with high accuracy of 92% AUC (NLST dataset), 86% AUC (MGH dataset) and 94% AUC (CGMH dataset) (Mikhael et al., 2023). Sybil is able to predict individual cancer risk without the need for demographic or clinical data, contributing to personalised screening approaches (Mikhael et al., 2023).

Similarly, Huang et al. (2023) compared machine learning and DL models to predict the survival of LC patients. In the study, the DNN model achieved the highest success with an accuracy of 88.58% compared to traditional logistic regression and decision trees. These results demonstrate the superiority of DL-based approaches in predicting survival time (Huang et al., 2023).

Despite these advancements, most of the existing studies concentrate on a single stage of the diagnostic pipeline—typically classification or segmentation—without integrating key components such as preprocessing, data balancing, architectural optimization, and comprehensive performance comparison. Moreover, data imbalance and generalization across multiple classes remain underexplored challenges. To address these issues, the present study proposes a unified DL-based classification pipeline for LC diagnosis using CT images. The approach includes comprehensive preprocessing, data balancing using the SMOTE, architectural tuning, and comparison of three robust CNN architectures—ResNet101, VGG19, and DenseNet121—on the IQ-OTH/NCCD dataset across three LC categories: benign, malignant, and normal.

3. MATERIAL AND METHOD

3.1. DataSet

The success of deep learning (DL) algorithms largely depends on access to large-scale, high-quality datasets. Sufficiently diverse and well-annotated datasets allow DL models to learn robust and generalizable features, reduce the risk of overfitting or underfitting, and ensure reliable performance across different patient populations. For this reason, standardized, ethically sourced, and clinically representative data are essential in developing accurate diagnostic systems.

In this study, we utilized the IQ-OTH/NCCD lung cancer dataset, a publicly available dataset hosted on the Kaggle platform (AL-Huseiny, 2021). This dataset was collected over three months in Fall 2019 from the Iraq Oncology Teaching Hospital and the National Center for Cancer Diseases. It comprises 1,190 axial CT scan slices from 110 individuals, including both healthy subjects and patients diagnosed with lung cancer at various stages. The cases were labelled and validated by experienced radiologists and oncologists.

The dataset includes three class categories: malignant, benign, and normal. Each CT scan contains between 80 to 200 slices, and images were originally acquired in DICOM format using a Siemens SOMATOM scanner. The imaging protocol included 120 kV tube voltage, 1 mm slice thickness, window width ranging from 350 to 1200 HU, and window center between 50 and 600 HU, with breath-hold at full inspiration. For compatibility with CNN models, the images were converted to JPEG format and resized to 224×224 pixels. The class distribution is shown in Table 1.

The dataset reflects real-world imaging variability and includes individuals from diverse geographic and socioeconomic backgrounds, particularly from central Iraq (e.g., Baghdad, Wasit, Diyala, Salahuddin, and Babylon).

Table 1. *Class Distribution of the IQ-OTH/NCCD Lung Cancer Dataset*

Class	Number of Images	Description
Normal	101	Healthy lung CT slices
Benign	432	Non-cancerous pulmonary nodules
Malignant	657	Cancerous lung tissue samples
Total	1190	

However, the publicly accessible version does not contain structured demographic information such as age, gender, or clinical history in machine-readable format. As a result, subgroup analyses based on patient-level metadata could not be conducted, which is acknowledged as a limitation of this study. All data were anonymized prior to release. The dataset's usage complies with ethical guidelines, and approval was granted by the institutional review boards of the participating medical centers. Written informed consent was waived as per local ethical protocols.

3.2. DL Algorithms

Deep Learning (DL) has emerged as a powerful tool in the field of healthcare, particularly in medical imaging and cancer diagnosis. By leveraging neural networks capable of learning complex patterns, DL enables automated analysis of radiological images, reducing the burden on clinicians and enhancing diagnostic accuracy. In oncology, DL models are widely used to detect and classify tumors from medical scans such as CT, MRI, and PET images. These models can identify subtle patterns that might be overlooked by human experts, leading to earlier and more precise diagnoses.

In cancer imaging, DL-based approaches are especially valuable for tasks such as tumor segmentation, classification, and progression monitoring. For lung cancer, convolutional CNN play a crucial role in detecting pulmonary nodules and distinguishing between benign and malignant lesions. By integrating DL into clinical workflows, healthcare professionals can improve diagnostic efficiency, personalize treatment plans, and ultimately enhance patient outcomes. As advancements continue, DL is expected to further revolutionize cancer imaging by offering real-time analysis, reducing diagnostic errors, and facilitating early intervention strategies.

In this study, three deep learning models ResNet101, VGG19, and DenseNet121 were selected for performance comparison in lung cancer classification. These models were chosen based on their proven effectiveness in medical image analysis, particularly in classification tasks involving CT and histopathological data. Additionally, these models offer architectural diversity in terms of depth and parameter complexity, enabling a comprehensive evaluation of model behaviour under the same training conditions. Their wide spread use in peer-reviewed literature ensures that the results can be reliably compared to existing studies and generalized to similar diagnostic contexts.

3.2.1. Denset121

DenseNet121 is a CNN architecture in which each layer creates dense connections by using the outputs from all previous layers. This structure prevents overlearning by increasing parameter efficiency and minimises the problem of gradient loss. Instead of connecting layers sequentially as in traditional CNNs, each layer receives input directly from all previous layers and transmits its output to all subsequent layers. DenseNet121 is widely used in areas that require high accuracy, such as medical image analysis, as it provides more efficient feature extraction with fewer parameters (Aslan, 2025). In a densely connected CNN, the input of each layer is expressed as follows:

$$x_1 = H_n([x_0, x_1, x_2, \dots, x_{n-1}]) \quad (1)$$

Where x_n is the output in n layers, H_n is the activation function and $x_0, x_1, x_2, \dots, x_{n-1}$ is the combination of the outputs of all previous layers.

3.2.2. ResNet50

ResNet101 (Residual Network) is a CNN architecture developed to address the problem of gradient loss in deep neural networks. Similar to ResNet50, it employs residual connections to facilitate learning and enable the construction of much deeper networks (Eren et al., 2024). ResNet101 is a 101-layer deep CNN architecture, widely used for tasks such as medical image classification, object detection, and segmentation. Due to its increased depth, it can capture more complex features, making it suitable for high-precision image analysis applications. It offers high accuracy and generalisation capacity, increasing the trainability of deeper networks. Instead of the standard linear transformation, ResNet blocks are modelled with residual connections as follows:

$$H(x) = F(x, W) + x \quad (2)$$

Here, $H(x)$ represents the activation value extracted from the block, while $F(x, W)$ denotes the conventional convolutional transformation. In addition, this transformation, computed using W weights, represents the filtering operations performed by the model on the input data x.

3.2.3. VGG19

VGG19 is a 19-layer CNN model with a deep but simple structure. The basic principle of the model is to create deeper feature maps and perform high accuracy classifications by using small (3x3) convolutional filters. Due to its simple yet effective architecture, it is widely preferred in tasks such as medical image analysis, object recognition and segmentation (Çelik & İnik, 2023). However, due to the large number of parameters, the computational cost is higher compared to other CNN architectures. In the VGG architecture, the active feature map in each convolutional layer is computed as follows:

$$y = \sigma(W * x + b) \tag{3}$$

Where W is the convolution kernel, "$*$" is the convolution process, b is the bias term and σ is the activation function. The VGG architecture performs a successful classification by extracting the features of the images at different scales with the max-pooling process following successive convolutional layers.

3.3. Evaluation of Performances

The evaluation of DL models in this study was conducted using various performance metrics. Additionally, confusion matrices, as well as model accuracy and error rates, were analyzed to provide a detailed insight into prediction performance. The accuracy score indicates how well the model identifies correct classifications across different categories, whereas precision measures the proportion of correctly predicted positive cases. Recall reflects the model's effectiveness in detecting true positive cases, while the F1 score balances these two measures, ensuring a comprehensive assessment of classification success.

The Confusion Matrix shows the correct and incorrect predictions made by the model, allowing the false positive (FP) and false negative (FN) rates to be analysed. This provides a detailed insight into which classes the model is more successful in and which errors it makes, particularly in LC diagnosis.

Model Accuracy and Model Error curves were analysed to understand the behaviour of the model during the training process. The Model Accuracy curve shows the change in the accuracy of the model during the training process, while the Model Error curve analyses the evolution of the model's error rates over time. These analyses helped to identify potential problems such as overfitting or underfitting. The results obtained provide a comprehensive performance analysis to assess how reliable the model is in LC diagnosis and whether it is suitable for clinical applications. Equations (4)-(7) allow the shortcomings of the model to be identified and areas for improvement to be identified for future studies.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$PRC = \frac{TP}{TP + FP} \tag{5}$$

$$RCL = SNS = \frac{TP}{TP + FN} \tag{6}$$

$$F1_{score} = \frac{2 * RCL * PRC}{RCL + PRC} \tag{7}$$

Accuracy (ACC), Precision (PRC), Recall (RCL) and Sensitivity (SNS).

3.4. Data Preprocessing, Balancing, and Model Optimization

Data preparation is essential for improving model performance in DL-based image classification tasks because it guarantees consistency and lowers noise in the input data. In this study, all CT images from the IQ-OTHNCCD LC dataset underwent multiple preprocessing steps before being fed into the DL model. These

steps include image resizing, grayscale conversion, noise reduction via Gaussian blur, and normalization. Each of these techniques contributes to improving model generalizability and robustness. Since the IQ-OTHNCCD LC dataset contains images of varying dimensions, it is essential to resize them to a uniform shape to maintain consistency in input size.

To ensure uniformity, all images were rescaled to 256×256 pixels through bilinear interpolation. This specific resolution was selected to optimize both processing efficiency and the retention of important details. A smaller resolution might lead to loss of crucial structural information, whereas a much higher resolution would increase computational costs without significant performance gains. Medical images often contain redundant colour information that does not contribute significantly to classification tasks. Converting CT images to grayscale helps remove unnecessary RGB colour channels, thereby reducing computational complexity and memory usage. More importantly, grayscale conversion enables the DL model to focus on essential structural patterns, such as edges, textures, and contrast variations that distinguish benign, malignant, and normal lung tissues.

Medical images frequently contain noise due to variations in scanning conditions, patient movement, or equipment differences. To enhance image quality and suppress unwanted artifacts, Gaussian blur was applied with a kernel size of (5,5). This technique smooths the image while preserving critical edges and texture structures. This technique smooths the image while preserving critical edges and texture structures. The result is a reduction in small, irrelevant variations that might otherwise lead to misclassification by the DL model. A normalization process was implemented to standardize pixel values across all images, adjusting them to fall within a 0 to 1 range.

One of the primary challenges in medical datasets is the imbalance between different classes, which can lead to biased model predictions. In the IQ-OTHNCCD LC dataset, there was a significant imbalance between benign, malignant, and normal cases. Without addressing this issue, the model might become skewed towards the majority class, thereby reducing its ability to accurately classify underrepresented classes.

To mitigate this issue, the SMOTE was employed. Unlike simple duplication of existing data points, SMOTE generates new synthetic examples for underrepresented classes, improving class balance. This method follows a structured approach: first, it randomly selects a minority class sample, then identifies its k-nearest neighbours, and finally, it synthesizes new data points by interpolating between the chosen instance and one of its neighbors. After SMOTE, all classes were balanced, resulting in an equal number of samples across categories. This adjustment allowed the DL model to learn from underrepresented cases effectively, thereby improving classification accuracy across all categories. Optimizing a DL model involves selecting appropriate architectural components, tuning hyperparameters, and employing regularization techniques to prevent overfitting. In this study, the ResNet101, VGG19 and DenseNet121 models were chosen as the base architecture, with several modifications to optimize its performance.

ResNet101 is a convolutional neural network that has been pre-trained on extensive image datasets. Rather than building the model from the ground up, transfer learning was employed to make use of features learned during prior training. Several modifications were introduced, including freezing convolutional layers and integrating custom fully connected layers. This approach allows for quicker model adaptation and improved generalization on the IQ-OTHNCCD LC dataset, as it utilizes previously extracted low-level features such as edges and textures while concentrating on learning more advanced patterns associated with lung cancer detection. To mitigate overfitting and avoid unnecessary computations, an early stopping strategy was implemented. This mechanism continuously evaluates validation loss throughout training and ceases further iterations if no improvement is detected within a predetermined number of epochs (set to 3 epochs). By doing so, the model prevents excessive training, which could otherwise result in poor generalization to new, unseen data. These techniques expanded the training dataset by introducing variations through transformations, including rotation, flipping, zooming, and contrast adjustments. Such alterations enhance the model's ability to recognize lung cancer patterns under different imaging conditions. formations such as: Rotation, Flipping, Zooming and Contrast Adjustment.

3.5. Training Strategy for the Proposed DL Model

The development process of the proposed model consists of four main stages: data preparation, data separation and normalisation, DL model construction, model training and model evaluation. In the first stage, images were processed using the IQ-OTHNCCD LC dataset, and this process is shown in Figure 1. During image resizing and preprocessing, all data were resized to 256×256 pixels and noise reduction methods were applied. In addition, the dataset was randomly shuffled by determining the class labels. In the second step, the pixel values of the images were normalised in the range of 0-1 and the dataset was divided into 75% training and 25% testing. To eliminate class imbalances, the SMOTE method was applied to ensure a balanced data distribution in the training set. In the third step, a ResNet101 based DL model was created, pre-trained layers were frozen and additional classification layers were added. The model was trained using the Adam optimization algorithm and a sparse categorical cross-entropy loss function. To improve training efficiency, an early stopping strategy was applied. Model predictions were validated against a separate dataset, and performance was evaluated using a confusion matrix and a classification report.

All model training, evaluation, and preprocessing operations were conducted using Python 3.8 in a Linux-based Google Colab Pro environment. The system was equipped with 2× NVIDIA Tesla T4 GPUs (16 GB each) and 25 GB RAM, which enabled efficient parallel computation during model training. The models were implemented using TensorFlow 2.x and the Keras API. Data handling and analysis were performed using Pandas, NumPy, and Scikit-learn 1.1.3. Image preprocessing utilized OpenCV, PIL, and imageio, while Matplotlib, Seaborn, and Plotly were used for visualization. Data balancing was applied using the Synthetic Minority Over-sampling Technique (SMOTE) from the imblearn library, and model training incorporated

early stopping and data augmentation using ImageDataGenerator. The experimental workflow was designed to be reproducible and aligned with best practices in deep learning-based medical image analysis.

This study aims to classify CT images using DL models for LC diagnosis. Advanced CNN architectures with powerful feature extraction capabilities such as ResNet101, VGG19 and DenseNet121 were used in the study. The dataset used is from the IQ-OTH/NCCD LC dataset and consists of 1,190 CT scan slices classified into three different classes as benign, malignant and normal. The main difference this study brings to the literature is that it provides a holistic approach to preprocessing, data imbalance removal and efficient optimisation of model architectures. In contrast to traditional methods, the SMOTE is used to address the class imbalance problem and the effect of the unbalanced distribution of the dataset on model performance is analysed. In addition to the final layers of the pre-trained models, additional optimisation layers were integrated to enable the model to better discriminate between different classes. In addition, early stopping mechanisms and data augmentation techniques were used in the training process to prevent overlearning and increase the generalisation capability of the model. The study comprehensively compares different DL architectures and analyses their accuracy rates, especially in the classification of malignant cases.

4. EXPERIMENTAL RESULTS

The experimental investigations carried out to assess the effectiveness of the suggested model and the outcomes are shown in this part. This study uses the IQ-OTHNCCD LC dataset to analyze the classification performance of various DL models. Data separation, model training, and evaluation procedures were applied methodically, beginning with preprocessing activities. Basic measures including accuracy, sensitivity, specificity, and F1 score were used to measure the model's performance and compare it to other studies.

Figure 2 shows examples of lung CT scans. The images are categorised as benign (a), malignant (b) and normal (c) to represent different lung conditions. The benign (a) image shows generally benign lesions or abnormal but non-cancerous tissue formations. A malignant (b) image may show marked irregularities in the lung tissue, indicating the presence of malignant cells. a normal (c) image shows no abnormal formation and the lung tissue appears to be healthy.

Figure 3 shows the confusion matrices generated to compare the performance of the DL models. Each matrix visualises the accuracy and error rates by showing the relationship between the predictions made by the model and the actual classifications. Figure 3a shows the confusion matrix for the ResNet101 model.

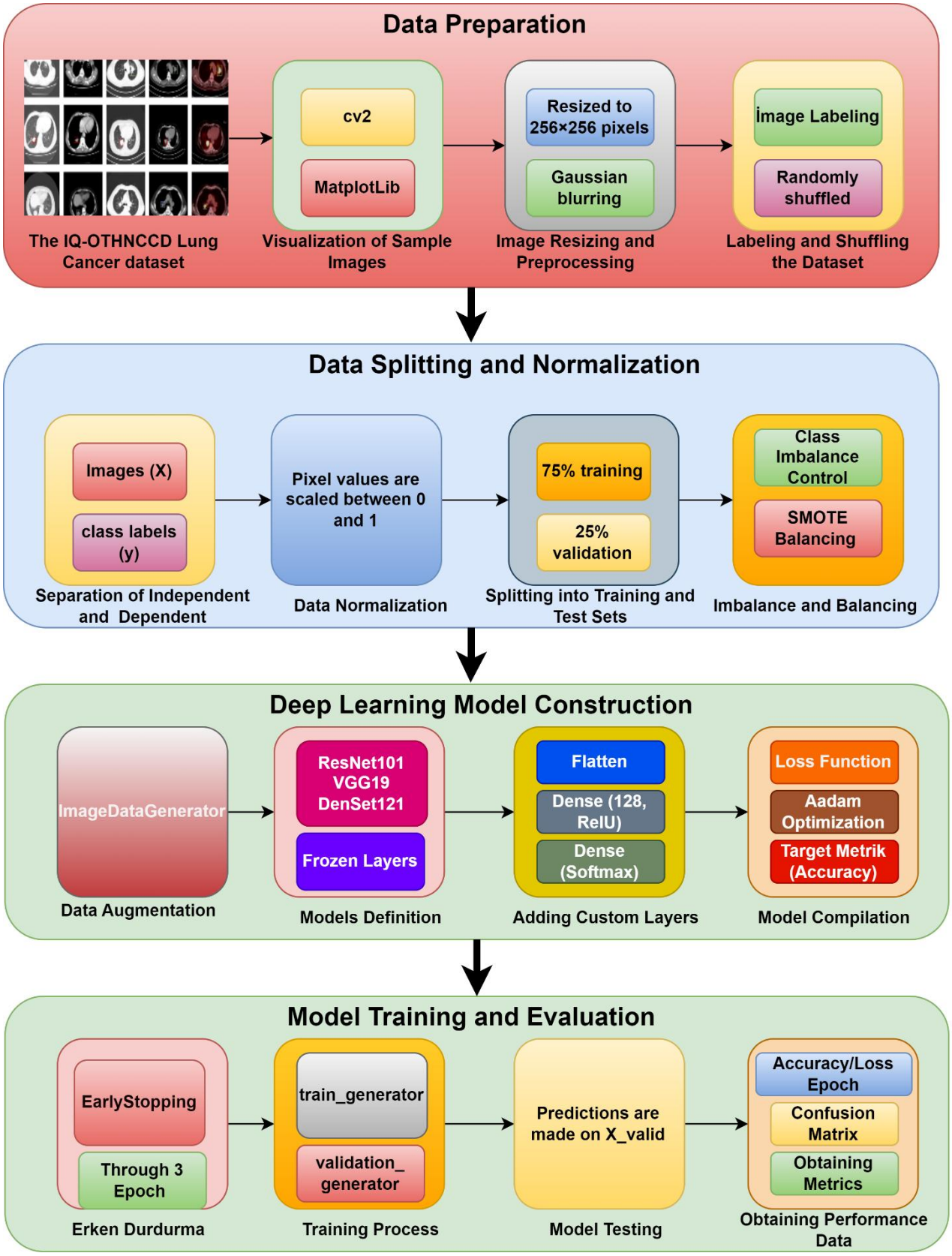
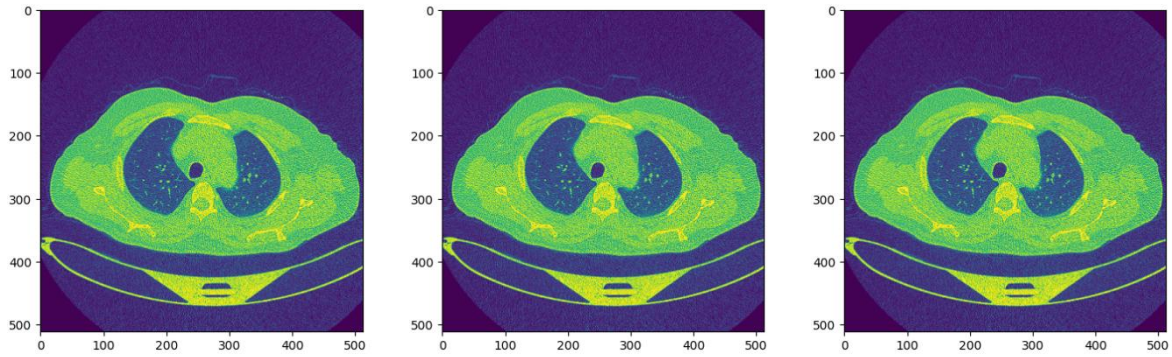


Figure 1. *Flowchart of data processing and model training for LC classification model*



a) b) c)

Figure 2. *Examples of lung CT images.* ***a)*** *Benign,* ***b)*** *Malignant,* ***c)*** *Normal*

The model classified malignant cases with 100% accuracy (143 correct predictions, 0 incorrect). However, 8 normal cases were misclassified as benign. This shows that the model is quite successful for the malignant class, but makes some mistakes in distinguishing between benign and normal classes. Figure 3b shows the performance of the VGG19 model. The model shows high accuracy for malignant cases, making only 1 error in this class. However, 7 normal cases were misclassified as benign. These results show that the model is successful in discriminating malignant cases, but is confused between benign and normal classes. Figure 3c shows the confusion matrix of the Dense121 model. Although the model correctly classified a large proportion of malignant cases (138 correct predictions, 4 errors), it misclassified more cases in the benign class than the other models. In particular, 9 benign cases were incorrectly predicted, indicating that the model performed worst in distinguishing between benign and normal classes.

The ResNet101 model stands out as the most successful model in terms of accurate prediction of malignant classes. However, incorrect prediction of normal cases as benign is a factor that can affect the overall performance of the model. While the VGG19 model has high accuracy in malignant class, it has limited performance in discriminating between normal and benign classes. The Dense121 model made more errors in predicting the benign class compared to the other models. The results show that each model has different classification abilities and that model selection should be made accordingly in applications focusing on a specific class.

Figure 4 shows the accuracy and loss values of the DL models during the training process. The performance of the ResNet101 (a), VGG19 (b) and DenseNet121 (c) models on training and validation sets are compared. The accuracy values show a steady increase as the training process progresses for all models. The fact that the training and validation accuracies are close to each other shows that the overall performance of the models is consistent. Analysing the loss values reveals that all models experience a notable decline at first, but that the subsequent epochs show a more balanced trend. Although the VGG19 model shows fluctuations in verification loss in some epochs, it generally shows a steady downward trend. The ResNet101 and DenseNet121 models, on the other hand, show more stable loss values and offer high performance in the validation set. When the results are evaluated, it can be seen that all models have successfully completed the training process. The

ResNet101 model achieved the highest accuracy value, while the DenseNet121 model showed a fast-learning process in the early epochs. The VGG19 model, on the other hand, shows a generally successful performance, although it fluctuates in verification accuracy from time to time. The results show that the models work consistently and effectively on different datasets.

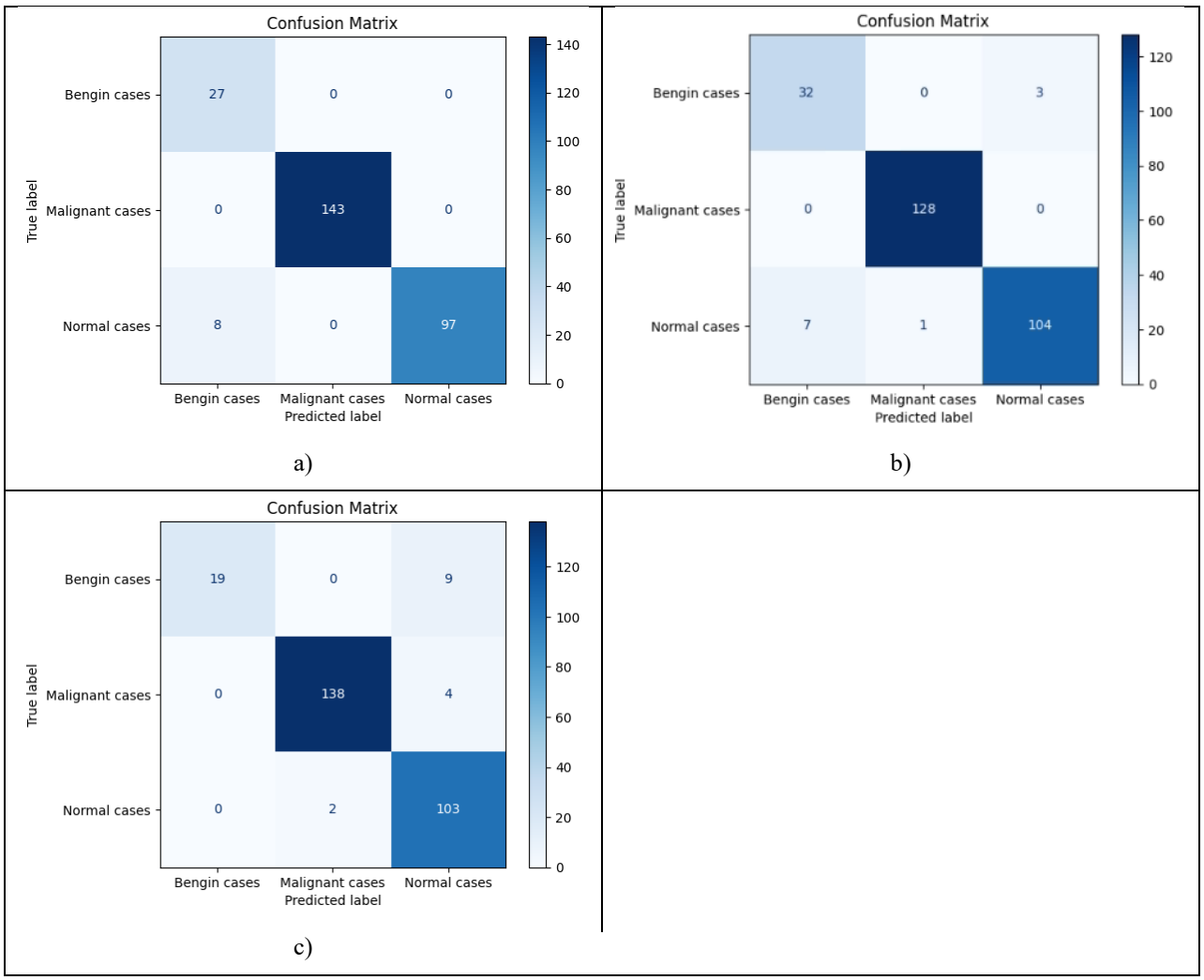


Figure 3. *Confusion Matrices for the DL model.* ***a)*** *ResNet101,* ***b)*** *VGG19,* ***c)*** *DenseNet121*

Figure 5 shows the ROC curve of the ResNet101 model. The curve almost reaches a true positive rate of 1 and maintains a false positive rate close to zero. The AUC (Area Under the Curve) value of 1.00 indicates that the model's classification ability is extremely high, with very few, if any, false positives or negatives. This result highlights the model's reliability, particularly in critical applications such as medical image classification, where accurate predictions are essential.

Figure 6 presents the Grad-CAM heatmaps for the normal, benign, and malignant cases, illustrating the model's focus on specific regions of the CT images for classification. In the normal case, the heatmap displays a relatively uniform distribution, highlighting the typical anatomical structures of the lungs without significant pathological changes. For the benign case, the heatmap indicates localized regions of interest, suggesting areas where benign abnormalities are detected. In contrast, the malignant case exhibits a more concentrated heatmap, with heightened intensity in certain regions, reflecting the model's identification of malignant features. These

results demonstrate how the Grad-CAM technique enables a visual interpretation of the model's attention, providing valuable insights into the decision-making process for classifying different types of pathologies in medical imaging.

Table 2 presents the classification performance of different DL models based on ACC, RCL/SNS, SPC, and F1-score.

Table 2. *Performance Comparison of DL Models*

Algorithm	ACC	RCL/SNS	SPC	F1 score
ResNet101	0.98	0.98	0.98	0.98
VGG19	0.96	0.96	0.96	0.96
DenseNet121	0.95	0.88	0.97	0.91

The ResNet101 model achieves the highest performance across all metrics, with an ACC of 0.98. The fact that RCL/SNS and SPC values are equal indicates that the model effectively balances positive and negative classifications. The VGG19 model also demonstrates strong performance, achieving an ACC of 0.96, along with equal RCL/SNS and SPC values, indicating stable classification performance. In contrast, the DenseNet121 model exhibits a relatively lower performance, particularly in terms of RCL/SNS, which is 0.88, despite showing competitive SPC (0.97) and F1-score (0.91). Overall, the ResNet101 model outperforms the other models, delivering the highest classification success. While the VGG19 model provides satisfactory accuracy and balanced sensitivity and specificity, the DenseNet121 model, despite strong SPC and F1-score values, falls behind in terms of sensitivity. These variations highlight critical performance differences that should be considered when selecting models for classification tasks. Table 3 presents a comparison between the classification performance of the proposed ResNet101 model and previous studies in the literature.

Table 3. *Comparison of performance of DL models with literature*

Algortim	ACC	RCL/SNS	SPC	F1 score
(Asuntha & Srinivasan, 2020)	0.95	0.97	0.96	-
(Shafi et al., 2022)	0.94	0.94	0.95	0.94
(Kumar et al., 2024)	0.95	-	-	0.85
(Mohamed & Ezugwu, 2024)	0.97	0.97	0.97	0.97
ResNet101(Proposed Model)	0.98	0.98	0.98	0.98

The evaluation, based on key metrics such as ACC, RCL/SNS, SPC, and F1-score, highlights that the proposed model achieves superior results compared to existing approaches. The ResNet101 model demonstrates the highest scores across all metrics, with 98% ACC, 98% RCL/SNS, 98% SPC, and a 98% F1-score. The highest ACC rate in the literature is 97%, as reported by (Mohamed & Ezugwu, 2024). However, although this study is at the same level as the proposed model in terms of RCL/SNS and SPC, it lags behind in terms of ACC.

Studies by (Asuntha & Srinivasan, 2020) and (Shafi et al., 2022) reported ACC values of 95% and 94%, respectively, showing lower performance than the proposed model. Furthermore, (Kumar et al., 2024)did not

report RCL/SNS and SPC values, and the F1-score was the lowest at 0.85%. In comparison, the proposed ResNet101 model proves to be more successful, exceeding the best results in the literature in terms of both ACC and other classification metrics.

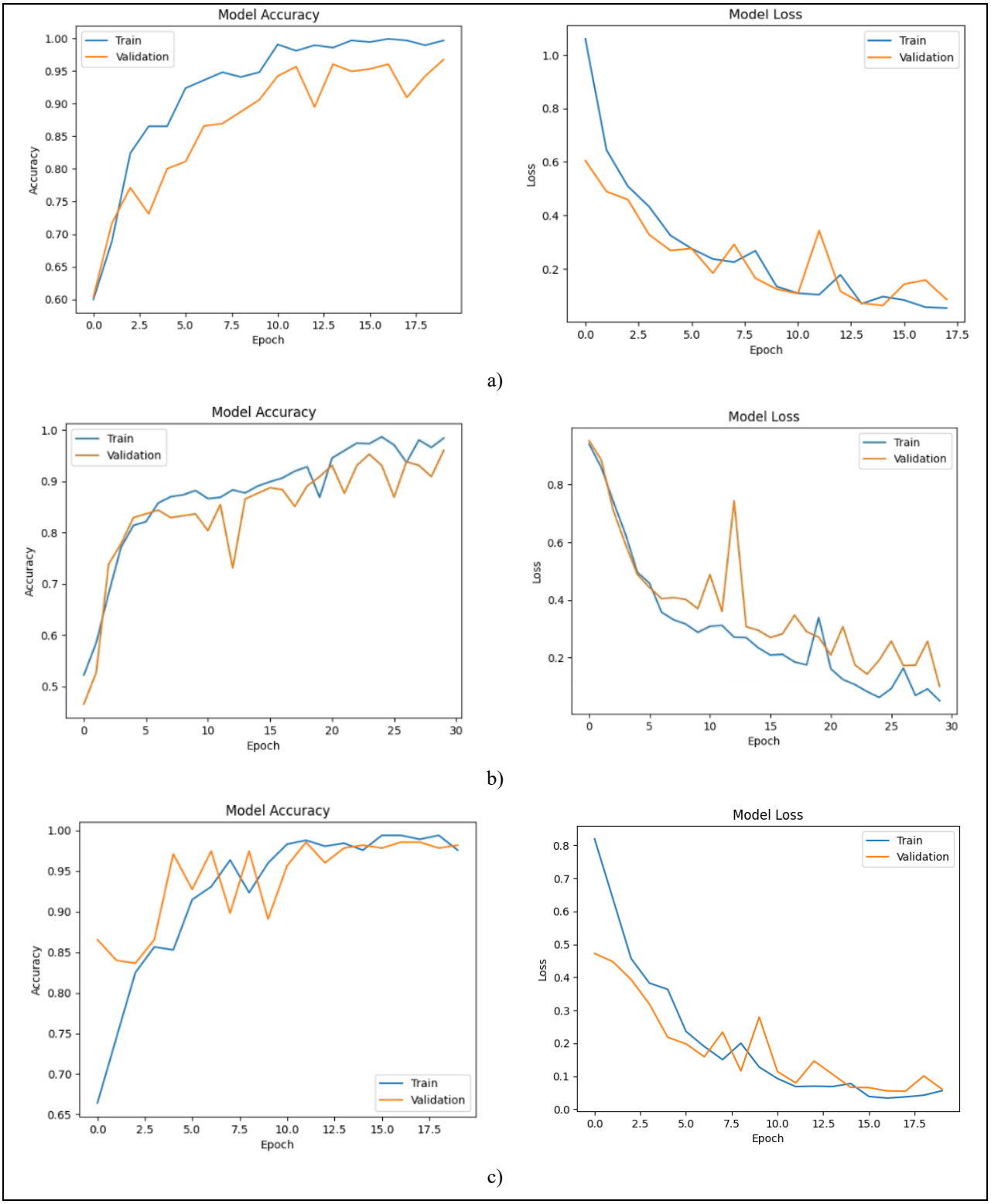


Figure 4. *Accuracy/Loss curves for the training process of the DL model.* ***a)*** *ResNet101,* ***b)*** *VGG19,* ***c)*** *DenseNet121*

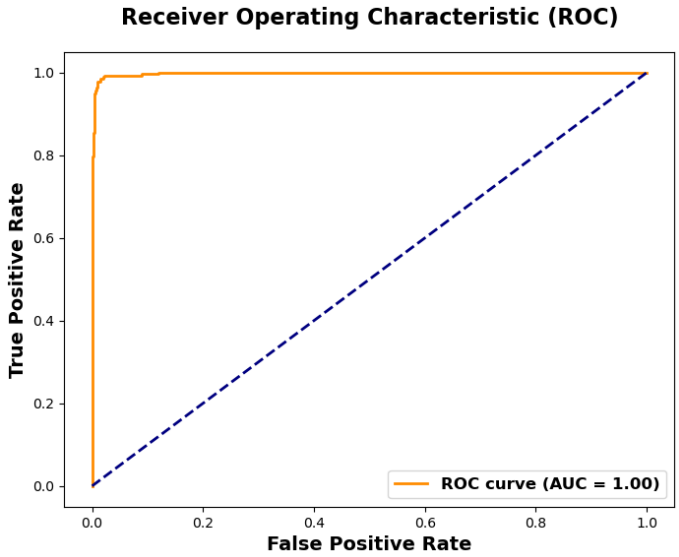


Figure 5. *ROC Curve of ResNet101 Model and Performance Evaluation*

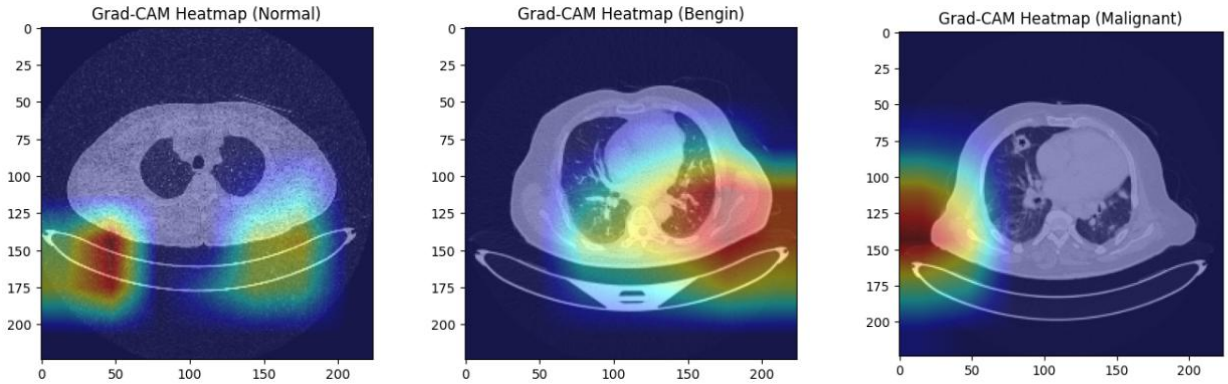


Figure 6. *Grad-CAM Heatmaps for Normal, Benign, and Malignant Cases*

This study demonstrated that the proposed ResNet101 model provided superior performance in classifying lung CT images, achieving the highest accuracy, sensitivity, specificity, and F1-score (all 98%) among the tested deep learning models (VGG19, DenseNet121) and existing studies in the literature. ROC analysis (AUC: 1.00) and Grad-CAM heatmaps further confirmed the model's reliability and enhanced interpretability by visualizing regions critical to classification. However, a limitation of the current study is the absence of sociodemographic information in the dataset, which may influence model generalizability.

5. CONCLUSION

This study investigated the effectiveness of DL models in LC diagnosis using CT images. Three CNN architectures—ResNet101, VGG19, and DenseNet121—were evaluated on the IQ-OTH/NCCD dataset to classify lung conditions as benign, malignant, or normal. To enhance model performance, class imbalance was addressed using SMOTE, data augmentation techniques were applied, and an early stopping mechanism was implemented to prevent overfitting.

The experimental results show that the ResNet101 model achieved the highest accuracy (98%) and outperformed other architectures in all key classification metrics. Notably, its superior performance in distinguishing malignant cases increases its potential for clinical applications. Comparisons with previous studies indicate that the proposed model offers superior accuracy and generalization capability. Despite these promising findings, some misclassifications occurred between benign and normal cases, suggesting that further improvements in feature extraction could enhance performance. Future research should explore hybrid architectures, multi-modal approaches, and the benefits of explainable AXI to improve interpretability and trust in clinical settings.

In conclusion, this study highlights the potential of DL-based models as a powerful tool for LC diagnosis, offering high accuracy, reliability, and efficiency. Integrating such models into clinical decision support systems could significantly improve early detection and patient outcomes.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

AL-Huseiny, M. (2021). National Cancer Center Database (IQ-OTH/NCCD - Lung Cancer Dataset) [https://www.kaggle.com/datasets/adityamahimkar/iqothnccd-lung-cancer-dataset/data]. In *Kaggle*. https://doi.org/doi: 10.17632/bhmdr45bh2.2

Aslan, E. (2025). Development of malaria diagnosis with convolutional neural network architectures: a CNN-based software for accurate cell image analysis. *ITEGAM-JETIA*, *11*(51), 35–42. https://doi.org/10.5935/JETIA.V11I51.1392

Asuntha, A., & Srinivasan, A. (2020). Deep learning for lung Cancer detection and classification. *Multimedia Tools and Applications*, *79*(11–12), 7731–7762. https://doi.org/10.1007/s11042-019-08394-3

Çelik, M., & İnik, Ö. (2023). Detection of Monkeypox Among Different Pox Diseases with Different Pre-Trained Deep Learning Models. *Iğdır Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, *13*(1), 10–21. https://doi.org/10.21597/JIST.1206453

Crasta, L. J., Neema, R., & Pais, A. R. (2024). A novel Deep Learning architecture for lung cancer detection and diagnosis from Computed Tomography image analysis. *Healthcare Analytics*, *5*(1) 100316. https://doi.org/10.1016/J.HEALTH.2024.100316

Davri, A., Birbas, E., Kanavos, T., Ntritsos, G., Giannakeas, N., Tzallas, A. T., & Batistatou, A. (2023). Deep Learning for Lung Cancer Diagnosis, Prognosis and Prediction Using Histological and Cytological Images: A Systematic Review. *Cancers 2023,* *15*(15), 3981. https://doi.org/10.3390/CANCERS15153981

Devarajan, H. R., Balasubramanian, S., Kumar Swarnkar, S., Kumar, P., & Jallepalli, V. R. (2023). Deep Learning for Automated Detection of Lung Cancer from Medical Imaging Data. International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI 2023).

https://doi.org/10.1109/ICAIIHI57871.2023.10488962

Eren, B., Fen, Ü., Dergisi, B., Aslan, E., & Özüpak, Y. (2024). Classification of Blood Cells with Convolutional Neural Network Model. *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, *13*(1), 314–326. https://doi.org/10.17798/BITLISFEN.1401294

Huang, S., Arpaci, I., Al-Emran, M., Kılıçarslan, S., & Al-Sharafi, M. A. (2023). A comparative analysis of classical machine learning and deep learning techniques for predicting lung cancer survivability. *Multimedia Tools and Applications*, *82*(22), 34183–34198. https://doi.org/10.1007/s11042-023-16349-y

Javed, R., Abbas, T., Khan, A. H., Daud, A., Bukhari, A., & Alharbey, R. (2024). Deep learning for lungs cancer detection: a review. *Artificial Intelligence Review, 57*(8), 1–39. https://doi.org/10.1007/S10462-024-10807-1

Kumar, V., Prabha, C., Sharma, P., Mittal, N., Askar, S. S., & Abouhawwash, M. (2024). Unified deep learning models for enhanced lung cancer prediction with ResNet-50–101 and EfficientNet-B3 using DICOM images. *BMC Medical Imaging*, *24*(1), 1–21. https://doi.org/10.1186/S12880-024-01241-4

Mamatha, B., Rashmi, D., Tiwari, K. S., Sikrant, P. A., Jovith, A. A., & Reddy, P. C. S. (2023). Lung Cancer Prediction from CT Images and using Deep Learning Techniques. 2023 2nd International Conference on Trends in Electrical, Electronics and Computer Engineering (TEECCON 2023), 263–267. https://doi.org/10.1109/TEECCON59234.2023.10335801

Mikhael, P. G., Wohlwend, J., Yala, A., Karstens, L., Xiang, J., Takigami, A. K., Bourgouin, P. P., Chan, P., Mrah, S., Amayri, W., Juan, Y. H., Yang, C. T., Wan, Y. L., Lin, G., Sequist, L. V., Fintelmann, F. J., & Barzilay, R. (2023). Sybil: A Validated Deep Learning Model to Predict Future Lung Cancer Risk from a Single Low-Dose Chest Computed Tomography. *Journal of Clinical Oncology*, *41*(12), 2191–2200. https://doi.org/10.1200/JCO.22.01345

Mohamed, T. I. A., & Ezugwu, A. E. S. (2024). Enhancing Lung Cancer Classification and Prediction With Deep Learning and Multi-Omics Data. *IEEE Access*, *12*(1), 59880–59892. https://doi.org/10.1109/ACCESS.2024.3394030

Mohamed, T. I. A., Oyelade, O. N., & Ezugwu, A. E. (2023). Automatic detection and classification of lung cancer CT scans based on deep learning and ebola optimization search algorithm. *PLOS ONE*, *18*(8), 0285796. https://doi.org/10.1371/JOURNAL.PONE.0285796

Özdemir, B., Aslan, E., & Pacal, I. (2025). Attention Enhanced InceptionNeXt Based Hybrid Deep Learning Model for Lung Cancer Detection. *IEEE Access*. https://doi.org/10.1109/ACCESS.2025.3539122

Said, Y., Alsheikhy, A. A., Shawly, T., & Lahza, H. (2023). Medical Images Segmentation for Lung Cancer Diagnosis Based on Deep Learning Architectures. *Diagnostics 2023, 13*(3), 546. https://doi.org/10.3390/DIAGNOSTICS13030546

Shafi, I., Din, S., Khan, A., Díez, I. D. L. T., Casanova, R. del J. P., Pifarre, K. T., & Ashraf, I. (2022). An Effective Method for Lung Cancer Diagnosis from CT Scan Using Deep Learning-Based Support Vector Network. *Cancers*, *14*(21), 5457. https://doi.org/10.3390/CANCERS14215457

Tárnoki, Á. D., Tárnoki, D. L., Dąbrowska, M., Knetki-Wróblewska, M., Frille, A., Stubbs, H., Blyth, K. G., & Juul, A. D. (2024). New developments in the imaging of lung cancer. *Breathe*, *20*(1), 230176. https://doi.org/10.1183/20734735.0176-2023

Thandra, K. C., Barsouk, A., Saginala, K., Aluru, J. S., & Barsouk, A. (2021). Epidemiology of lung cancer. *Contemporary Oncology*, *25*(1), 45. https://doi.org/10.5114/WO.2021.103829

Tran, T. O., Hoa Vo, T., & Khanh Le, N. Q. (2024). Omics-based deep learning approaches for lung cancer decision-making and therapeutics development. *Briefings in Functional Genomics*, *23*(3), 181–192. https://doi.org/10.1093/BFGP/ELAD031

Wang, L. (2022). Deep Learning Techniques to Diagnose Lung Cancer. *Cancers, 14*(22), 5569. https://doi.org/10.3390/CANCERS14225569

Wani, N. A., Kumar, R., & Bedi, J. (2024). DeepXplainer: An interpretable deep learning based approach for lung cancer detection using explainable artificial intelligence. *Computer Methods and Programs in Biomedicine*, *243*(1*)*, 107879. https://doi.org/10.1016/J.CMPB.2023.107879

Zhang, Y., Yang, Z., Chen, R., Zhu, Y., Liu, L., Dong, J., Zhang, Z., Sun, X., Ying, J., Lin, D., Yang, L., & Zhou, M. (2024). Histopathology images-based deep learning prediction of prognosis and therapeutic response in small cell lung cancer. *Npj Digital Medicine 2024*, *7*(1), 1–12. https://doi.org/10.1038/s41746-024-01003-0