Advancing Diagnostic Accuracy in Lung Disease Severity Classification Using Multi-Domain Features

Urvashi Deshmukh, Prapti Deshmukh

***Abstract:** Accurate classification of lung diseases is crucial for early diagnosis and effective treatment. This study presents an optimised classification framework that utilises multi-domain feature extraction and a deep neural network (DNN) for categorising lung disease severity from CT scan images. The dataset collected from a local hospital includes 266 CT scans of Lung cancer, COVID-19, and pneumonia, categorized into mild (43 images), moderate (82 photos), and severe (141 images) cases. To address class imbalance, the synthetic minority oversampling technique (SMOTE) was applied, ensuring equal representation across categories. A total of 30 multi-domain features were extracted using a comprehensive feature extraction methodology that combined wavelet packet decomposition (WPD) with statistical, texture, shape, edge detection and Grey Level Co-occurrence Matrix (GLCM) features. These features captured diverse spatial and frequency-based characteristics of lung disease patterns, enabling robust model input. This study focuses on classification based on the severity of patient condition within three different classes: Mild, Moderate, and Severe, related to lung disease. The classification was performed using a Deep Neural Network (DNN) with fine-tuned hyperparameters. The model achieved a training accuracy of 95%. The findings underline the potential of this approach in improving automated diagnostic systems. The extracted features provide a comprehensive representation of disease patterns, while the DNN leverages these features for precise classification. This methodology offers valuable insights for applications in medical imaging. This research contributes to the field of medical image analysis by integrating robust feature extraction techniques with advanced classification models, paving the way for more accurate and reliable lung disease diagnosis.*

***Keywords:** Synthetic Minority Oversampling Technique, Wavelet Packet Decomposition, Grey Level Co-occurrence Matrix, Deep Neural Network.*

Abbreviations:
DNN: Deep Neural Network
SMOTE: Synthetic Minority Oversampling Technique
WPD: Wavelet Packet Decomposition
CAD: Computer-Aided Diagnosis
GLCM: Grey Level Co-occurrence Matrix
METS: Metabolic Syndrome
SMOTE: Synthetic Minority Oversampling Technique
RLGLN: Run-Length Grey Level Nonuniformity
NTM: Neighbour Texture Mean
ROC: Receiver Operating Characteristics
TP: True Positives
FP: False Positives
TN: True Negatives
FN: False Negative
SVM: Support Vector Machine
KNN: K-Nearest Neighbours

Manuscript received on 26 March 2025 | First Revised Manuscript received on 06 April 2025 | Second Revised Manuscript received on 17 June 2025 | Manuscript Accepted on 15 July 2025 | Manuscript published on 30 July 2025.

*Correspondence Author (s)

Urvashi Deshmukh*, Scholar, Department of Computer Science, Dr. G. Y. Pathrikar College of CS and IT, MGM University, Aurangabad (Maharashtra), India. Email ID: urvashi.deshmukh17@gmail.com, ORCID ID: 0009-0002-3374-9547

Prapti Deshmukh, Department of Computer Science, Dr. G. Y. Pathrikar College of CS and IT, MGM University, Aurangabad (Maharashtra), India. Email ID: prapti.research1@gmail.com

I. INTRODUCTION

Lung diseases, including lung cancer, pneumonia, and COVID-19, remain significant global health challenges. Early diagnosis and precise classification of disease severity are crucial for effective treatment and management. Computer-aided diagnosis (CAD) systems, utilising advanced classification algorithms, have emerged as pivotal tools in this domain. These systems utilise machine learning and deep learning techniques to analyse CT scans, providing consistent and reliable diagnoses. Recent studies have demonstrated that combining robust feature extraction [1] with optimized classification methods enhances the accuracy of automated systems. For instance, [2] emphasized that integrating texture and statistical features improved the differentiation between mild, moderate, and severe cases. Furthermore, the adoption of deep learning models [3], particularly convolutional neural networks, has shown promise in outperforming traditional classifiers by learning intricate patterns in medical images [4]. Despite advancements, existing classification models often suffer from limitations such as overfitting, sensitivity to imbalanced datasets, and inability to capture multi-domain features comprehensively. These challenges underscore the need for frameworks that integrate diverse feature sets, including spatial, frequency, and texture-based attributes, to represent the complexity of lung diseases accurately. Moreover, the limited dataset sizes in medical imaging studies exacerbate these challenges, underscoring the importance of utilising techniques such as synthetic data generation and balanced datasets to ensure model robustness. This study aims to develop and evaluate an optimized classification framework for lung disease severity using multi-domain features extracted from CT scans. By combining Wavelet Packet Decomposition (WPD), statistical, textural, shape, edge detection, and GLCM features, we seek to construct a comprehensive feature set that enhances classification accuracy. A Deep Neural Network (DNN) model is employed to evaluate the classification performance.

A comprehensive review of research papers highlights the progress and gaps in lung disease classification techniques (Sawant & Sreemathy, 2022) [1]. This paper reviews various texture feature extraction techniques applied to chest CT images for the detection and classification of pulmonary diseases. It highlights traditional and advanced texture-based descriptors and discusses their effectiveness across diagnostic systems.

Dharmalingam & Kumar (2022) [2] The authors propose a hybrid feature selection model combining statistical and heuristic methods to enhance lung disorder classification. The study emphasizes increased accuracy and robustness of machine learning classifiers through optimal feature selection.

Goyal & Singh (2023) [3] This study compares machine learning and deep learning approaches for classifying pneumonia and COVID-19 from chest images. It concludes that a deep understanding, particularly with CNNs, achieves higher precision and better generalisation performance.

Bourouis et al. (2020) [4] This review explores hybrid approaches for medical image analysis, focusing on combining traditional image processing with intelligent algorithms. It outlines current challenges and suggests future directions for improving diagnostic accuracy.

Ali et al. (2024) [5] The authors introduce a wavelet transform-based deep learning framework for lung cancer detection. Their method leverages multi-resolution analysis and CNN architectures to improve feature representation and classification results.

Karthika, Rajaguru, & Nair (2024) [6] This study introduces a statistical framework combining wavelet feature extraction and bio-inspired optimisation for lung cancer prognosis. The approach enhances classification performance by selecting the most discriminative features from CT images.

Saihood, Karshenas, & Nilchi (2022) [7] The authors propose a deep fusion technique using gray level co-occurrence matrices (GLCM) for lung nodule classification. The model improves diagnostic precision by integrating multiple texture descriptors into a deep learning framework.

Koshta et al. (2024) [8] This paper presents a Fourier decomposition-based method for the automated classification of healthy, COPD, and asthma conditions using single-channel lung sounds. The system demonstrates high accuracy, offering a non-invasive diagnostic solution.

Abdulazeez, Zeebaree, & Abdulqader (2020) [9] A comprehensive review of wavelet transform applications in medical imaging is provided, covering its roles in noise reduction, compression, and feature extraction. The paper highlights the use of wavelet transforms in enhancing image analysis tasks.

Serte, Dirik, & Al-Turjman (2022) [10] The study evaluates several deep learning models for detecting COVID-19 from chest X-ray images. It highlights model performance in real-world datasets and discusses challenges in deploying AI-based detection tools in clinical settings.

Mahmood & Ahmed (2022) [11] This paper presents an enhanced CNN architecture designed for automatic classification of lung nodules. The proposed model outperforms existing architectures by improving accuracy and reducing false positives, showcasing its clinical relevance in early lung cancer detection.

Kaur, Goyal, & Dogra (2023) [12] The authors develop a hybrid feature-based model combining handcrafted and deep learning features for computer-aided diagnosis of lung cancer. Their approach improves the robustness and accuracy of lung cancer detection systems using CT images.

Piffer et al. (2024) [13] This systematic review addresses the challenge of limited data in medical image classification and explores AI-based strategies to overcome it. Techniques like data augmentation, transfer learning, and synthetic data generation are discussed in depth.

Chamseddine et al. (2022) [14] The paper focuses on managing class imbalance in COVID-19 chest X-ray datasets using SMOTE and weighted loss functions. These techniques significantly enhance the performance of deep learning models in detecting underrepresented cases of COVID-19.

Koetzier et al. (2024) [15] This paper discusses the generation of synthetic medical imaging data to augment training datasets for deep learning models. It emphasizes techniques like GANs and highlights their potential to address data scarcity while ensuring patient privacy.

Ali et al. (2021) [16] The authors propose a method combining deep feature selection with decision-level fusion for lung nodule classification. Their approach improves diagnostic performance by integrating diverse features and optimizing classifier decisions.

Krstinić et al. (2020) [17] This work explores performance evaluation for multi-label classifiers using confusion matrices. It provides a structured framework for assessing classification metrics, especially relevant for medical imaging tasks involving multiple disease labels.

Carrington et al. (2021) [18] The study introduces Deep ROC analysis and advocates for AUC as a balanced average accuracy metric. It offers insights for improving model selection and interpretability in deep learning applications within the healthcare sector.

Al-qaness et al. (2024) [19] This comprehensive survey reviews deep learning methods applied to chest X-ray images for detecting various lung diseases. It covers a wide range of architectures, preprocessing techniques, and challenges in real-world implementation.

Karalis (2024) [20] The paper discusses how artificial intelligence is being integrated into clinical practice, highlighting its benefits, limitations, and ethical considerations. It emphasizes the importance of aligning AI tools with clinical workflows for effective adoption.

These studies collectively highlight the potential of combining diverse feature extraction techniques with optimised classification models to enhance the diagnosis of lung diseases.

The paper is structured as follows: Section 2 discusses the materials and methods, including the dataset, feature extraction techniques, classification models and evaluation matrices. Section 3 presents the results, focusing on model performance metrics and visualisations. Section 4 provides a detailed discussion, including insights, challenges, and comparisons with existing techniques. Section 5- concludes the study, summarizing the findings and suggesting directions for future work.

II. METHODOLOGY

A. Dataset Description

The dataset for this study comprises CT scan images of 266 subjects diagnosed with lung cancer, COVID-19, and pneumonia collected from MGM Medical College and Hospital, Aurangabad, Maharashtra.

The CT images in the dataset were categorised into three severity levels based on diagnostic hypotheses, which were verified with doctors. The Mild category consisted of 43 images, where a single mass was observed in the scan. The Moderate category included 82 images, characterized by the presence of a mass with nodular involvement. The Severe category comprised 141 images, where the mass was accompanied by metabolic syndrome (METS).

Due to the dataset's imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied to generate additional samples, thereby ensuring an equal distribution across all three categories. As a result, the dataset was expanded to a total of 426 images, with each severity level having the same number of samples, thereby improving the robustness of the classification model.

B. Preprocessing Steps

The preprocessing steps involved several techniques to standardise and enhance the quality of the images before they were used for classification. Resizing was performed to ensure that all photos had uniform dimensions, providing a consistent input size for the model. Normalization was applied to adjust the intensity values, ensuring uniform contrast and brightness across the dataset, which helped in improving feature extraction and model performance. To further enhance image quality, noise reduction was implemented using a Gaussian filter, which effectively minimized noise while preserving essential image details.

Following preprocessing, the dataset was divided into three subsets: 70% for training, 15% for validation, and 15% for testing, ensuring a structured approach to model evaluation. The balanced dataset, obtained after applying augmentation techniques, was then used for multiclass classification of Mild, Moderate, and Severe cases. This prepared dataset served as the input for the classification models, facilitating a robust and reliable assessment of lung disease severity.

C. Features Used for Classification

In this study, a total of 30 features were extracted from lung disease images to analyze texture, shape, and edge characteristics. These features are divided into two main categories: Wavelet Packet Decomposition (WPD) features and complementary features. The Wavelet Packet Decomposition (WPD) feature consists of 12 distinct attributes that capture detailed frequency and spatial information. These features are derived from various decomposition pathways, including horizontal, vertical, diagonal, and approximation directions at multiple levels. By breaking down lung tissue images into finer components, WPD features provide insights into subtle variations in texture. These features are crucial for detecting slight changes in lung tissue, rendering them highly valuable for disease classification. Complementary Features consist of 18 attributes, which are further divided into statistical, texture, shape, edge detection, and Grey Level Co-occurrence Matrix (GLCM) features. Statistical Features such as skewness and kurtosis describe the distribution of pixel intensities. Skewness quantifies asymmetry in pixel values, while kurtosis measures the peakedness of the intensity distribution. Skewness is mathematically defined as:

$$Skewness(X) = \frac{\sum (x_i - \mu)^3}{n.\sigma^3} \quad \dots \quad (1)$$

where μ The mean and standard deviation (σ) are used to describe the data.
Kurtosis given by:

$$Kurtosis(X) = \frac{\sum (x_i - \mu)^4}{n \cdot \sigma^4} - 3 \quad \dots \quad (2)$$

Texture features include Run-Length Grey Level Nonuniformity (RLGLN) and Neighbour Texture Mean (NTM). RLGLN highlights non-uniformity in intensity distribution along image gradients, expressed as:

$$RLGLN(X) = \sum \left(r_{i,j} - \mu_r\right)^2 \quad \dots \quad (3)$$

Where $r_{i,j}$represents the run length and μ_r It is the average run length. NTM, on the other hand, represents the mean intensity of neighbouring pixels and is given by:

$$NTM(X) = \frac{\sum x_{i,j}}{n \times m} \quad \dots \quad (4)$$

Shape features, such as compactness, centroid, and perimeter, focus on the geometric properties of lung tissues. Compactness is calculated as:

$$Compactness = \frac{4\pi \times A}{P^2} \quad \dots \quad (5)$$

Where A is the area and P is the perimeter. These shape features provide a structural understanding of lung abnormalities.

Edge detection features, such as Canny and Sobel edges, identify boundaries between tissue regions by detecting gradients and transitions. Canny edge detection works by computing gradients using the convolution of the image with Gaussian filters, and the Sobel operator uses approximations of derivatives to find edges. Finally, GLCM features, such as correlation and dissimilarity, analyse the relationships between neighbouring pixels to quantify texture patterns. Correlation measures the linear relationship between pixel pairs, while dissimilarity quantifies differences in intensity values between pairs. Together, these features offer a comprehensive framework for the detailed analysis and classification of lung tissue abnormalities, enhancing the model's ability to distinguish between different severities of lung diseases.

III. CLASSIFICATION MODEL

The deep neural network (DNN) model presented in this study is designed to classify lung disease severity into three distinct categories: Mild, Moderate, and Severe.

By leveraging advanced deep learning techniques, including dense layers, activation functions, regularisation methods, and optimization strategies, the model aims to achieve high accuracy and robustness in classification.

A. Model Architecture

The model begins with an input layer, which receives data with a dimensionality defined by X_train, corresponding to the number of features extracted from the lung images. This input layer serves as the initial representation of the features before they are passed through multiple dense layers for feature extraction and refinement. The hidden layers of the model consist of multiple densely connected layers, each progressively refining the feature extraction process. The first hidden layer contains 256 neurons with a ReLU activation function, capturing complex patterns from the input data. To enhance stability and convergence, batch normalization is applied to normalize the layer outputs. Additionally, a dropout rate of 30% is introduced to regularize the model, preventing overfitting by randomly deactivating neurons during training.

The second hidden layer reduces the complexity while maintaining significant feature representations by employing 128 neurons with ReLU activation. Similar to the first layer, batch normalization is applied for output stabilization, and dropout is used to enhance generalization. The third hidden layer further refines feature extraction with 64 neurons, maintaining the same regularization techniques to ensure the model remains stable and avoids overfitting. The fourth and final dense hidden layer consists of 32 neurons, serving as the last stage of feature refinement before classification.

The output layer of the model comprises three neurons, each corresponding to one of the three categories of lung disease severity. A soft max activation function is employed, ensuring that the output represents probability distributions across the three classes. The softmax function converts raw model outputs into probabilities, where the sum of all class probabilities equals 1.

i. Activation Functions

Activation functions play a crucial role in learning complex patterns. The model primarily employs the Rectified Linear Unit (ReLU) activation function in the hidden layers, which is defined as:

$$ReLU(x) = (0, x) \quad \ldots \quad (6)$$

ReLU introduces non-linearity, enabling the model to learn intricate patterns while mitigating the vanishing gradient problem. For the output layer, the softmax activation function is utilized, which is mathematically defined as:

$$P(x) = \frac{e^{z_k}}{\sum_{j=1}^{C} e^{z_j}} \quad \ldots \quad (7)$$

Where z_k represents the linear output for the class k. This function ensures that the predicted probabilities sum to one, facilitating multi-class classification.

ii. Loss Function and Optimization

For training, the model utilises sparse categorical cross-entropy as the loss function, which is particularly suitable for multi-class classification. It is given by:

$$L = -\frac{1}{N}\sum_{i=1}^{N} log\; log\; (p_{i,y_i}) \quad \ldots \quad (8)$$

Where p_{i,y_i} represents the predicted probability for the actual class y_i of the i^{th} sample. This loss function effectively quantifies the difference between predicted and actual class probabilities.

To optimise the learning process, the model utilises the Adam optimizer, which combines the advantages of momentum and RMS Prop for adaptive weight updates. The update rule for Adam is:

$$\widehat{\theta_{t+1}} = \theta_t - \frac{\eta \cdot m_t}{\sqrt{v_t} + \epsilon} \quad \ldots \quad (9)$$

Where m_t and v_t The moving averages of the first and second moments of the gradients, and η Represents the dynamically adjusted learning rate. This optimization technique ensures efficient convergence while adapting the learning rate based on gradient information.

iii. Regularization Techniques

To enhance generalization and mitigate overfitting, the model incorporates batch normalisation and dropout as regularisation strategies. Batch normalization is applied after each dense layer to normalise outputs and stabilise the learning process, reducing internal covariate shifts. Dropout is implemented with a 30% dropout rate, which randomly deactivates neurons during training, preventing over-reliance on specific features and improving model robustness.

iv. Learning Rate Adjustment

The learning rate is dynamically adjusted during training using Reduce LR on Plateau, which reduces the learning rate by a factor of 0.5 when the validation loss stagnates for 10 consecutive epochs. Additionally, Early Stopping is employed to halt training if the validation loss does not improve for 20 straight epochs, thereby preventing unnecessary computations and minimising the risk of overfitting.

v. Training Parameters

The model is trained for a maximum of 2000 epochs, ensuring sufficient iterations for convergence while leveraging early stopping to terminate training when performance plateaus. A batch size of 32 is used to maintain a balance between computational efficiency and effective model convergence.

B. Evaluation Metrics

In the context of evaluating a classification model for lung disease severity, several key metrics are utilized to assess the model's performance. These metrics include accuracy, precision, recall, F1-score, and the receiver operating characteristic (ROC) curve. Accuracy measures the proportion of correctly predicted instances out of the total cases. It provides a simple and intuitive understanding of how well the model performs. However, it may not capture performance equally well across all classes, especially when dealing with imbalanced datasets. Precision is defined as the ratio of accurate optimistic predictions to the sum of accurate positive and false optimistic predictions. It reflects how accurately the model identifies positive instances. Recall (sensitivity) measures the model's ability to identify all positive cases correctly. It is the ratio of true positives to the sum of true positives and false negatives. High recall indicates the model's ability to capture as many positive cases as possible, which is crucial in medical diagnostics. F1-score is the harmonic mean of precision and recall, providing a balanced metric where both precision and recall contribute equally. It is beneficial in scenarios where there is an uneven class distribution. ROC Curve plots the actual positive rate (sensitivity) against the false positive rate (specificity) at various thresholds, offering a graphical representation of the model's performance across different classifications. The area under the ROC Curve (AUC-ROC) quantifies the model's ability to distinguish between classes. These metrics, when used together, provide a comprehensive evaluation of the model's performance in classifying lung disease severity, offering insights into both individual metrics and their combined impact on the model's overall effectiveness.

i. Confusion Matrix

The confusion matrix is a vital tool for understanding the performance of a classification model, providing a breakdown of true positives, false positives, true negatives, and false negatives.

True Positives (TP) represent the number of correctly predicted instances of the positive class. For lung disease classification, this corresponds to correctly identifying severe cases of lung disease. False positives (FP) are instances that are incorrectly predicted as positive, meaning non-severe cases of lung disease. True Negatives (TN) indicate correct prediction of the negative class, i.e., correctly identifying mild or moderate cases. False Negative (FN) represent cases where severe instances are missed by the model, indicating a model's inability to capture positive cases.

For these values, metrics like precision, recall, and F1-score are derived. The confusion matrix provides a comprehensive view of how the model handles different classes, thus enabling detailed performance evaluation. For example, a high number of false negatives indicates a lack of sensitivity, while a high number of false positives could indicate low precision. Analysing the confusion matrix helps in adjusting thresholds and improving model performance, ensuring a more accurate classification of lung disease severity levels.

IV. RESULT

A. Evaluation of Model Performance

To evaluate the performance of the classification model, which is developed to assess lung disease severity across three datasets: training, validation, and testing. The data was split into 70% for training, 15% for validation, and 15% for testing, ensuring that the model was trained on a substantial sample while leaving sufficient data for unbiased validation and testing. The model's performance is evaluated using key metrics such as accuracy, precision, recall and F1-score, along with an analysis of the confusion Matrix. These matrices provide a comprehensive understanding of the model's ability to classify disease severity into three categories: Mild, Moderate, and Severe. The confusion matrix was analysed to extract True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) rates for each dataset, enabling a deeper understanding of the model's behaviour.

i. Training set Performance

The training set results demonstrate the model's strong ability to classify the severity of lung disease accurately. With an overall accuracy of approximately 95%, the model shows near-perfect classification across all three classes. Precision, recall and F1-scores for each class are consistently high, reflecting the model's strong performance.

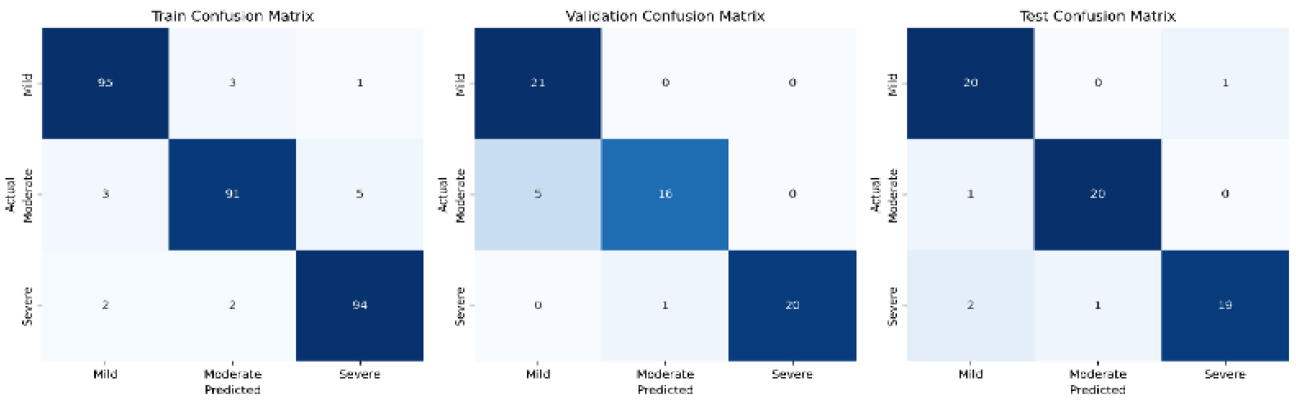
The confusion matrix for the training set reveals minimal misclassifications. For instance, Mild cases were rarely misclassified as Moderate or severe and the same trend was observed for other classes. The balance between precision (0.94-0.97) and recall (0.94-0.96) indicates that the model effectively identifies actual positive cases while minimizing false positives and false negatives. This strong performance in the training set establishes a solid foundation for generalization to unseen data.

ii. Validation Set Performance

The validation set results provide an unbiased evaluation of the model's ability to generalize to new data. The model achieved an accuracy of approximately 90%, which is slightly lower than the accuracy of the training set, as expected. Precision and recall remained high for mild and severe classes, but moderate showed a slight decrease in recall, which affected its F1-score. The confusion matrix for the validation set highlights some misclassification in the moderate category, where a few moderate cases were identified as mild. This misclassification explains the reduced recall for the moderate class. Despite these challenges, the severe class maintained excellent performance, with precision and recall exceeding 0.95. The validation results demonstrate the model's ability to generalise effectively, with only minor performance trade-offs.

iii. Test set Performance

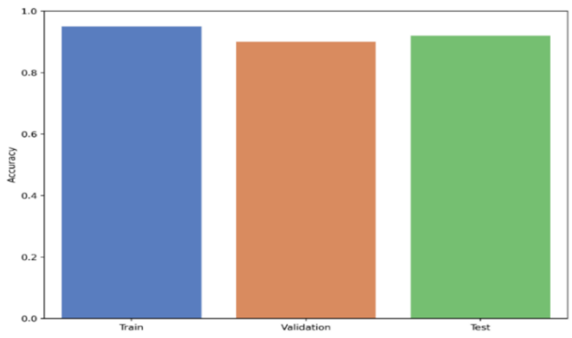
The test set results confirm the model's robustness and ability to handle unseen data. An overall accuracy of approximately 92% was achieved, closely matching the validation performance. Precision and recall for severe cases were consistently high, showcasing the model's ability to identify critical cases of lung disease severity with confidence. The confusion matrix for the test set indicates minimal misclassification. A few mild cases were misclassified as severe, and some moderate cases were misclassified as mild. However, these misclassifications were infrequent and did not significantly impact overall performance. The precision, recall and F1-score for all classes remained strong, particularly for severe cases, which achieved an F1-score of 0.90.



[Fig.1: Confusion Matrix Across Training, Validation, and Testing Datasets]

B. Accuracy, Precision, Recall, and F1-Score Analysis

Accuracy is a key metric that summarises the proportion of correctly classified instances across all classes. The training set achieved the highest accuracy of 95%, reflecting the model's strong performance on the data it was trained on. The validation set accuracy dropped slightly to 90%, while the test set achieved an accuracy of 92%. These results indicate a good balance between fitting the training data and generalising to new datasets.



[Fig.2: Accuracy Comparison Across Training, Validation, and Testing Datasets]

Precision, recall, and F1-score provide a detailed view of the model's performance in distinguishing between classes. These metrics are vital for imbalanced datasets or scenarios where certain misclassifications carry more significant consequences (e.g., misclassification of severe cases).

i. Training Set

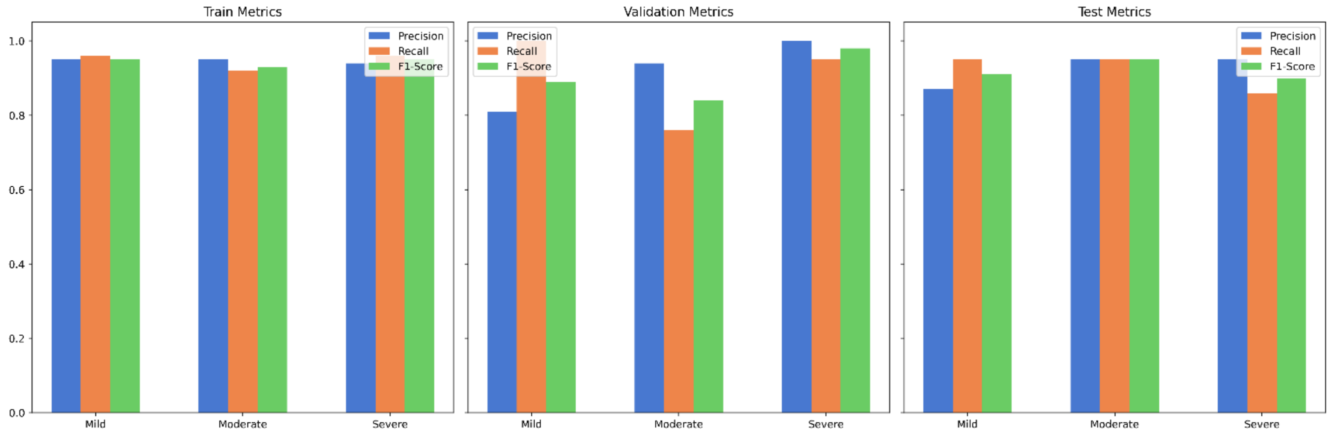
All three metrics were consistently high across classes, with precision ranging from 0.94 to 0.97 and recall from 0.94 to 0.96. The F1 scores demonstrated the model's balanced performance in identifying actual positive cases while minimising false positives and false negatives.

ii. Validation Set

Precision and recall remained high for Mild and severe classes, while the moderate class experienced a slight drop in recall, affecting its F1-score. This reflects a slight trade-off in performance when generalising to unseen data.

iii. Test Set

Results were comparable to those of the validation set, with a slight improvement in the precision of the moderate class. Severe cases consistently performed well across all metrics, highlighting the model's reliability in detecting critical cases.



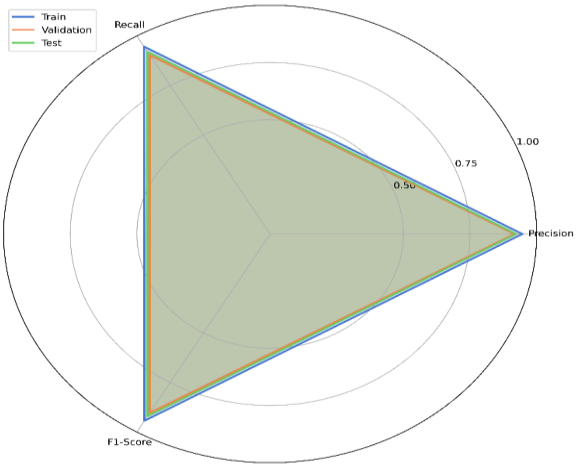
[Fig.3: Precision, Recall, and F1-Score Comparison Across Datasets]

C. Radar Chart Analysis

A radar chart offers a visual representation of the model's performance across precision, recall, and F1-score for each class and dataset.

The training set chart shows a uniformly high matrix for all classes, emphasizing the model's strong performance on the training data. In the validation set, the moderate class exhibits a slight decrease in recall, while the other classes maintain high values across all metrics.

The Test set results closely. Resemble the training set, with only minor variations, demonstrating the model's consistency across the dataset.



[Fig 4: Radar Chart of Precision, Recall, and F1-Score for All Classes]

D. ROC Curve Analysis

The ROC curve evaluates the model's ability to distinguish between classes, with the Area Under the Curve (AUROC) quantifying this ability. Higher AUROC values indicate better discriminatory power.

i. *Training Set*

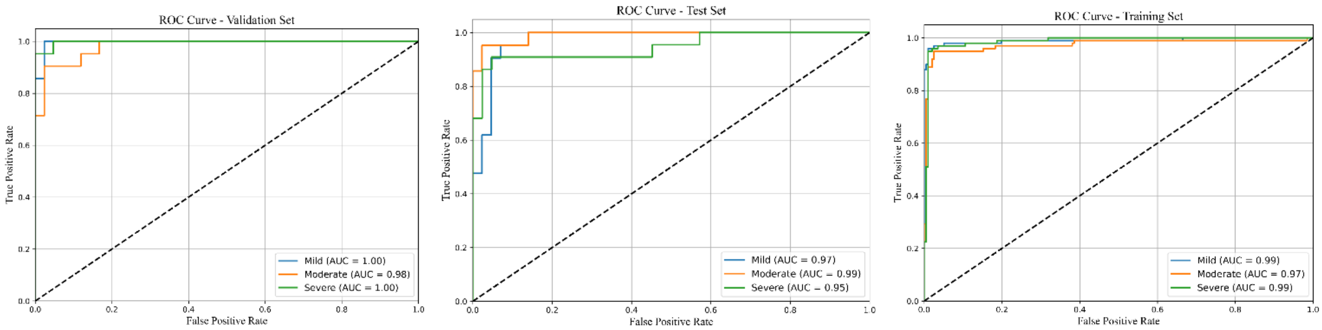
AUROC values were near perfect, with mild (0.99), moderate (0.97), and severe (0.99), reflecting excellent performance

ii. *Validation Set*

Mild achieved a perfect AUROC (1.00), while moderate and severe maintained high values of 0.98 and 1.00, respectively.

iii. *Test Set*

AUROC values were slightly lower but still strong-mild (0.97), moderate (0.99) and severe (0.95). The high AUROC values across all datasets highlight the model's effectiveness in distinguishing between severity levels, even under varying conditions.



[Fig 5: ROC Curve Comparison for Training, Validation, and Testing Datasets]

In addition to evaluating our proposed model, we conducted a comparative analysis with several traditional machine learning algorithms, including decision Trees, Random Forests, support vector machines (SVM), naïve Bayes, and K-nearest neighbours (KNN). During this analysis, we observed that many of these models exhibited signs of overfitting, especially when trained on the same dataset. This overfitting tendency suggests that these traditional algorithms struggled to generalise effectively to unseen data, further highlighting the robustness and superior generalisation capability of our approach.

V. DISCUSSION

The results of the lung disease severity classification model demonstrate robust performance across training, validation and testing datasets. Emphasizing its capability to generalize well to unseen data. Despite the relatively small dataset used, the model consistently achieved high accuracy, precision, recall, and F1-scores, confirming its reliability in distinguishing between Mild, Moderate, and severe cases.

The training set, achieving 95% accuracy, reflects the model's strong learning from the available data, with uniformly high performance across all metrics. This indicates a well-optimized model that effectively minimizes misclassification while accurately identifying the actual disease severity levels. The validation set, with an accuracy of 90%, highlights the model's ability to generalize. A slight decrease in recall for the Moderate class affected its F1-score, but overall performance remained robust, with Severe cases demonstrating excellent classification. The test set confirmed the model's reliability, achieving 92% accuracy, with substantial precision and recall, particularly for Severe cases, which consistently scored above 90%.

One of the model's key strengths is its ability to maintain high performance despite the small dataset. This is particularly noteworthy in medical applications, where acquiring large, annotated datasets is often a challenging task. These high AUROC values across all datasets, exceeding 0.95 for most classes, further underscore the model's strong discriminative capability. These results suggest that the model is effective not only in learning from the training data but also in distinguishing between severity levels in real-world scenarios.

However, the Moderate class exhibited slightly lower recall during validation and testing, indicating room for improvement in identifying these cases. This could be addressed by incorporating data augmentation or rebalancing techniques to mitigate potential class imbalance. Despite this, the model consistently excelled.

In classifying severe cases, which is crucial for clinical decision-making, it is essential to ensure that the most urgent cases are accurately identified and prioritised.

Overall, the model demonstrates a reliable and efficient approach to classifying the severity of lung disease. Its performance across key matrices, coupled with its ability to generalize well, highlights its potential as a valuable tool in aiding clinical diagnosis and treatment planning. Further refinement through advanced techniques could enhance its capabilities, particularly for challenging cases, making it an even more robust solution for medical imaging analysis.

A. Strengths of the Model

The following are the strengths, limitations, and implications of the model's performance –

High Overall Accuracy: Across all datasets (training, validation, and test), the model achieved high overall accuracy, with values consistently above 90%. Specifically, the training set attained an accuracy of approximately 95%, the validation set around 90%, and the test set approximately 92%. This demonstrates the model's robustness in distinguishing between different levels of lung disease severity, even with slight variations in data distribution across datasets.

Balanced Precision and Recall: The classification reports indicate that the precision and recall values were balanced across Mild, Moderate, and Severe classes. For instance, the F1-scores were above 0.90 in most cases, highlighting the model's ability to maintain a good balance between capturing positive cases (recall) and minimizing false positives (precision). Particularly in the Severe class, the model demonstrated excellent performance, with both precision and recall approaching 1.0, as evident in the confusion matrices for all sets.

Consistent Performance for Severe Cases: The Severe class consistently demonstrated strong results across all sets, with high precision, recall, and F-1 scores. This indicates that the model is particularly effective at accurately identifying cases of severe lung disease, which is crucial for medical diagnosis and treatment.

Effectiveness of SMOTE in addressing Class Imbalance: The dataset was augmented using the Synthetic Minority Oversampling Technique (SMOTE), which contributed to improved model performance by ensuring that each category had a balanced number of samples. This approach helped reduce the impact of class imbalance, resulting in more reliable predictions.

High Performance with Small Datasets: Another significant strength of the model is its ability to achieve high performance even with relatively small datasets. Despite having limited data, the model produced impressive results, demonstrating its capability to extract meaningful features and perform well in classifying lung disease severity levels.

B. Limitations and Areas for Improvement

i. Performance Drop in Validation Set

While the training set showed the highest accuracy and balanced metrics, the validation set displayed a slight reduction in performance, especially for Moderate cases. The precision and recall for the Moderate class were lower, reflecting the model's challenge in accurately distinguishing Moderate cases. Further optimization, such as fine-tuning the classification thresholds or incorporating additional features, could help improve this aspect.

Impact of False Positives and False Negatives: Although the overall performance was strong, the confusion matrices indicate the presence of both false positives (FP) and false negatives (FN). For instance, in the training set, Moderate and Severe classes showed higher false positives, indicating potential misclassification in complex regions of lung tissue. Similarly, false negatives highlight instances where severe cases were misclassified into less severe categories, emphasizing the need for more refined feature extraction and model interpretation.

Generalisation Across Diverse Datasets: Although the model performed well across the training and test sets, the slight decrease in performance for the validation set suggests a need for improved generalisation capabilities. This issue could be addressed by ensuring that the validation set is more representative of real-world data or by employing techniques such as cross-validation for a more comprehensive evaluation.

Implications for Medical Diagnostics: The model's ability to distinguish between different severity levels of lung disease has a significant impact on medical diagnostics. By improving the precision and recall, particularly for Moderate cases, healthcare professionals can benefit from accurate disease classification, leading to better management and treatment outcomes for patients. Additionally, the use of machine learning techniques such as SMOTE helps enhance the model's ability to handle imbalanced datasets, ensuring a more equitable diagnostic tool. In conclusion, while the lung disease severity classification model demonstrates strong performance, ongoing refinement and optimisation are necessary to address its limitations, particularly in distinguishing Between Moderate cases. Continuous advancements in feature extraction methods and model evaluation will enhance the model's ability to deliver accurate and reliable diagnostic insights across varying degrees of lung disease severity.

VI. CONCLUSION

The classification model developed for assessing lung disease severity demonstrates robust and reliable performance across training, validation, and testing datasets. With consistently high accuracy, precision, recall, and F1 Scores, the model effectively distinguishes between Mild, Moderate, and severe cases, making it a valuable tool for medical diagnostics. The use of advanced techniques, such as SMOTE, for handling class imbalance and efficient feature extraction, significantly contributed to its success, particularly in achieving high performance despite the challenges posed by relatively small datasets. The model's generalisation capability was validated through its strong performance on unseen data, as evidenced by its stable metrics across both the validation and testing datasets. In contrast, the traditional machine learning algorithms, such as Decision Trees, Random Forest, SVM, Naive Bayes, and KNN, were found to be prone to overfitting, further underscoring the efficacy of the proposed approach. However, a slight drop in recall for Moderate cases highlights an area for improvement. Future work could involve fine-tuning hyperparameters, incorporating data augmentation, and exploring ensemble techniques to enhance the model's ability to classify Moderate cases accurately while maintaining its strengths in identifying Severe cases. Overall, the proposed model provides a reliable and efficient solution for classifying lung disease severity, with significant potential to support clinical decision-making and enhance outcomes. With further refinements, it could serve as a robust tool in medical imaging analysis, paving the way for more accurate and accessible healthcare technologies.

DECLARATION STATEMENT

After aggregating input from all authors, I must verify the accuracy of the following information as the article's author.

- **Conflicts of Interest/Competing Interests:** Based on my understanding, this article does not have any conflicts of interest.
- **Funding Support:** This article has not been sponsored or funded by any organization or agency. The independence of this research is a crucial factor in affirming its impartiality, as it was conducted without any external influence.
- **Ethical Approval and Consent to Participate:** The data provided in this article is exempt from the requirement for ethical approval or participant consent.
- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible.
- **Author's Contributions:** The authorship of this article is contributed equally to all participating individuals.

REFERENCES

1. Sawant, P., & Sreemathy, R. (2022, December). A Review of Texture Feature Analysis in Chest Computed Tomography Images for the Detection and Classification of Pulmonary Diseases. In International Conference on Communication and Intelligent Systems (pp. 463-475). Springer Nature Singapore. DOI: https://doi.org/10.1007/978-981-99-2100-3_36
2. Dharmalingam, V., & Kumar, D. (2022). A hybrid feature selection model for the classification of lung disorders. Journal of Ambient Intelligence and Humanized Computing, 13(12), 5609-5625. DOI: https://doi.org/10.1007/s12652-021-03224-7
3. Goyal, S., & Singh, R. (2023). Detection and classification of lung diseases for pneumonia and COVID-19 using machine and deep learning techniques. Journal of Ambient Intelligence and Humanized Computing, 14(4), 3239-3259. DOI: https://doi.org/10.1007/s12652-021-03464-7
4. Bourouis, S., Alroobaea, R., Rubaiee, S., & Ahmed, A. (2020). Toward effective medical image analysis using hybrid approaches—Review, challenges, and applications. Information, 11(3), 155. DOI: https://doi.org/10.3390/info11030155
5. Ali, N. T., El Abbadi, N. K., & Ghandour, A. M. (2024). Lung cancer detection using wavelet transform with deep learning algorithms. BIO Web of Conferences, 97, 00050. DOI: https://doi.org/10.1051/bioconf/20249700050
6. Karthika, M. S., Rajaguru, H., & Nair, A. R. (2024). Wavelet feature extraction and bio-inspired feature selection for the prognosis of lung cancer—A statistical framework analysis. Measurement, 238, 115330. DOI: https://doi.org/10.2139/ssrn.4697036
7. Saihood, A., Karshenas, H., & Nilchi, A. R. N. (2022). Deep fusion of grey-level co-occurrence matrices for lung nodule classification. PLOS One, 17(9), e0274516. DOI: https://doi.org/10.1371/journal.pone.0274516
8. Koshta, V., Singh, B. K., Behera, A. K., & TG, R. (2024). Fourier decomposition-based automated classification of healthy, COPD, and asthma using single-channel lung sounds. IEEE Transactions on Medical Robotics and Bionics.] DOI: https://doi.org/10.1109/TMRB.2024.3408325
9. Abdulazeez, A. M., Zeebaree, D. Q., & Abdulqader, D. M. (2020). Wavelet applications in medical images: A review. Transform. DWT, 21,22.URL: https://www.researchgate.net/publication/341977072_Wavelet_Applications_in_Medical_Images_A_Review
10. Serte, S., Dirik, M. A., & Al-Turjman, F. (2022). Deep learning models for COVID-19 detection. Sustainability, 14(10), 5820. DOI: https://doi.org/10.3390/su14105820
11. Mahmood, S. A., & Ahmed, H. A. (2022). An improved CNN-based architecture for automatic lung nodule classification. Medical & Biological Engineering & Computing, 60(7), 1977-1986. DOI: https://doi.org/10.1007/s11517-022-02578-0
12. Kaur, B., Goyal, B., & Dogra, A. (2023, March). A hybrid feature-based model development for computer-aided diagnosis of lung cancer. In 2023 10th International Conference on Computing for Sustainable Global Development (INDIA Com) (pp. 1031-1036). URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10112386&tag=1
13. Piffer, S., Ubaldi, L., Tangaro, S., Retico, A., & Talamonti, C. (2024). Tackling the small data problem in medical image classification with artificial intelligence: A systematic review. Progress in Biomedical Engineering. DOI: https://doi.org/10.1088/2516-1091/ad525b
14. Chamseddine, E., Mansouri, N., Soui, M., & Abed, M. (2022). Handling class imbalance in COVID-19 chest X-ray image classification: Using SMOTE and weighted loss. Applied Soft Computing, 129, 109588. DOI: https://doi.org/10.1016/j.asoc.2022.109588
15. Koetzier, L. R., Wu, J., Mastrodicasa, D., Lutz, A., Chung, M., Koszek, W. A., ... & Willemink, M. J. (2024). Generating synthetic data for medical imaging. Radiology, 312(3), e232471. DOI: https://doi.org/10.1148/radiol.232471
16. Ali, I., Muzammil, M., Haq, I. U., Amir, M., & Abdullah, S. (2021). Deep feature selection and decision-level fusion for lung nodule classification. IEEE Access, 9, 18962-18973. DOI: https://doi.org/10.1109/ACCESS.2021.3054735
17. Krstinić, D., Braović, M., Šerić, L., & Božić-Štulić, D. (2020). Multi-label classifier performance evaluation with a confusion matrix. Computer Science & Information Technology, 1, 1-14. DOI: https://doi.org/10.5121/csit.2020.100801
18. Carrington, A. M., Manuel, D. G., Fieguth, P. W., Ramsay, T., Osmani, V., Wernly, B., ... & Holzinger, A. (2021). Deep ROC analysis and AUC as balanced average accuracy to improve model selection, understanding, and interpretation. arXiv preprint arXiv:2103.11357. DOI: https://doi.org/10.1109/TPAMI.2022.3145392
19. Al-qaness, M. A., Zhu, J., Al-Alimi, D., Dahou, A., Alsamhi, S. H., Abd Elaziz, M., & Ewees, A. A. (2024). Chest X-ray images for lung disease detection using deep learning techniques: A comprehensive survey. Archives of Computational Methods in Engineering, 31(6), 3267-3301. DOI: https://doi.org/10.1007/s11831-024-10081-y
20. Karalis, V. D. (2024). The integration of artificial intelligence into clinical practice. Applied Biosciences, 3(1), 14-44. DOI: https://doi.org/10.3390/applbiosci3010002

AUTHOR'S PROFILE

Ms. Urvashi B. Deshmukh is currently pursuing research as a Research Scholar at Dr. G. Y. Pathrikar College of Computer Science and IT, MGM University, located in Chhatrapati Sambhaji Nagar, India. The focus of their doctoral research lies in the advanced and rapidly evolving domain of Medical Image Processing, which integrates artificial intelligence techniques to enhance diagnostic accuracy in healthcare. Hold a Master of Computer Applications (Engineering) degree from the prestigious Savitribai Phule Pune University, Pune, India. With a strong academic background and technical proficiency, they have accumulated over 8 years of teaching experience in the field of Computer Science and Information Technology.

Dr. Prapti Deshmukh holds a Ph.D. in Computer Science, along with dual postgraduate degrees—M.Sc in Computer Science and M.Sc. in Physics—from Dr. Babasaheb Ambedkar Marathwada University, Chhatrapati Sambhaji Nagar, India. With a distinguished academic and research background, Dr. Prapti Deshmukh currently serves as a Research Guide and the Dean of the Faculty of Basic and Applied Sciences at MGM University, Chhatrapati Sambhaji Nagar. Additionally, they hold the esteemed position of Principal at Dr. G. Y. Pathrikar College of Computer Science and IT, where they have been instrumental in fostering academic excellence and research-driven education. Active member of several renowned scientific and professional bodies, including being a member of the IEEE, Indian Science Congress Association, Kolkata. Research expertise in Biometrics, Image Processing, IOT, GIS, Robotics, Pattern Recognition, Computer Vision, and Neural Network.

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