

Lung Disease Detection with Explainable AI and Deep Learning: A Comprehensive Study

Kavita Kolpe¹, MehzabinPathan², Pranit Sarode³, Sejal Rokade⁴, Pranav Sakpal⁵ and Avanti Sangai⁶

¹⁻⁶Pimpri Chinchwad College of Engineering, Department of Computer Engineering Pune, India

Email: kavita.kolpe@pccoepune.org, mehzabinpathan@pccoepune.org, pranit.sarode22@pccoepune.org, sejal.rokade22@pccoepune.org, pranav.sakpal22@pccoepune.org, avanti.sangai22@pccoepune.org

Abstract—Lung diseases, which include COVID-19, pneumonia, and tuberculosis, are significant global public health concerns and have the potential for serious health outcomes if not identified in time. This work proposes an explainable AI model that uses deep learning techniques to identify and predict different types of lung diseases from medical image data, such as chest X-rays and CT scans. This shall further allow a proper classification model by using convolutional neural networks to classify properly and make the proper diagnosis of the input based on the medical image while using the techniques of explanation like Grad-CAM, SHAP, and LIME to make the prediction understandable, thereby developing confidence of healthcare professionals in applying such models into the medical practices. It thus explores improvement of diagnostics with this improved model against classic issues facing the domain, data imbalance, and demand for explainability in AI-driven diagnosis. Results of this study show improvement in the detection of lung diseases while emphasizing the importance of transparency, which could result in improved patient outcomes and diagnostic reliability.

Index Terms— Deep Learning, Explainable AI, Lung Diseases, Convolutional Neural Networks (CNN), Medical Imaging, LIME, SHAP, Diagnostic techniques.

I. INTRODUCTION

The importance of research on deep learning in detecting and classifying lung diseases has increased because of the growing burden of respiratory disorders in the world. AI provides possibilities for early and accurate diagnosis. The last few years have seen a significant advancement in deep learning models and architectures- CNNs, transfer learning, and hybrid models, which have improved the feature extraction capability and, in turn, the classification of lung diseases such as COVID-19, pneumonia, and tuberculosis. Most explainable-AI methods, including LIME, SHAP, and Grad-CAM, have been extremely well publicized to enhance model transparency. Therefore, such AI models may gain more physicians' trust through better understanding. However, the model also has generalizability challenges and faces problems related to data imbalance as well as data used being too homogeneous in many ways. These issues highlight the continuing imperative of further research to create reliable, interpretable, and scalable systems that can seamlessly find utility in different clinical setups.

II. LITERATURE SURVEY

Advancements in AI have resulted in the development of the deep learning model, which may lead to delivering

accurate and timely diagnosis with the increase in the incidence of respiratory disorders worldwide. The field has become very important in research about the detection and classification of lung disease using deep learning. Feature extraction and classification are now being tapped through various deep learning models and architectures focusing on COVID-19, pneumonia, and tuberculosis. The more the model is becoming transparent, the more methods for Explainable AI such as LIME, SHAP, and Grad-CAM emerge due to the increasing awareness of physicians and consequently the confidence built in the prediction done using AI. These models are found to be accurate up to good but still have problems with data imbalance, limited diversity of the dataset, and generalization. These gaps show a need for more research in better solid, interpretable, and scalable solutions applicable across clinical platforms.

Irtaza et al. 2024 published multi-label classification studies for lung disease through deep models that are MobileNet, DenseNet, and VGG-16 over the primary NIH Chest X-ray 14 data. GANs have been employed for the purpose of augmenting data besides class balancing also, however here too, balancing performance by using GAN was not very high-value because synthetic data that was being generated have had a minor diversity as the reason. Classification outcomes from the multilabel lung forward by mobile Net produced very astounding accuracy values up to 93.4%, respectively [1].

Kieu et al. (2020) reviewed more than 98 studies published between the years 2016 to 2020, mainly aspects that include types of images feature extraction techniques and deep learning methods with CNNs, transfer learning, and ensemble classifiers. The study offered a taxonomy of lung disease detection through deep architectures with CNNs, specially transfer learning, which improves accuracy. This study, however, has generalization issues and major over-reliance on a particular lung disease [2].

Pitroda et al. (2021), propose a designed CNN architecture suitable for the lung disease classification problem using Layer-wise Relevance Propagation, Deep Taylor Decomposition and LIME for interpretability. It achieved an accuracy of 84% on DenseNet-201 but produced explanations indicating unconnected irrelevant regions such as bone structures; hence the paper calls out for better complex bone suppression methods in clinical trust [3].

Xie et al. (2024) suggested a lung disease classification system from chest X-ray images based on CNNs, hybrid models, ensemble approaches, hyperparameter tuning, stratified k-fold cross-validation, and transfer learning. The 5-fold cross-validation accuracy was about 92.21% for the Xception model. However, the poor diversity in the dataset and data augmentation dependency affected the model's generalization ability [4].

A two-phased experiment was conducted where PFTs were interpreted by pulmonologists, one with and one without XAI support that came out from the study conducted by Das et al. (2024). Phase 1 showed the improvement of diagnostic accuracy by 10.4% while Phase 2 showed an improvement of 5.4%, but the authors concluded that further studies in large datasets must be carried out to support clinical utility with XAI [5].

This methodology in medical images under XAI taxonomy comes under three key heads that are the visual, textual, and example-based explanation developed by van der Velden et al. in 2024. Although clinician's confidence is being boosted up by this visual explanation making Grad-CAM, this would not match all the requirements of real-time application demands so it has been forwarded by non-visual explanations which visual explanations would lack compared to them [6].

Mahamud et al. (2024) applied the fine-tuned transfer learning that uses DenseNet201 along with the XAI methods, such as SHAP, LIME, Grad-CAM, and Grad-CAM++, to improve lung disease classification interpretability. This model increased the clarity of the clinicians but had class imbalances with data augmentation being necessary due to the underlying problem of the given dataset [7].

Islam et al. (2024) integrated feature extraction with DCNN and temporal analysis with the help of a GRU for the classification of lung abnormality. XAI techniques such as LIME, SHAP, Grad-CAM were added to the model in order to make the outputs of the model interpretable. The model was tested on just two datasets, which does not generalize its utility; however, other datasets could prove the feasibility of the model in different clinical conditions [8].

In 2020, Bharati et al. had proposed a hybrid model called VDSNet, merging CNN, VGG16 and STN into one for getting an enhanced feature with better identification towards lung disease. Although this made the features come with an improved value using STN, it suffered from an extra computational cost which was associated along with slow convergence during training epochs that were limited. The images turned out to be large, requiring more training [9].

In pre-processing approaches of image quality improvement and extraction of lung-specific regions of interest from an image, using SVM, ANN, and ensemble classifiers, Goyal and Singh in 2023 determined pneumonia and COVID-19 in the images. An excellent classification accuracy was achieved. However, in achieving such good accuracy, such an extensive feature set also poses potential challenges in terms of scaling this up for more deployment [10].

Sultana et al. (2023) discussed the use of CNN models, including EfficientNetB0, DenseNet169, and DenseNet201, for the classification of COVID-19, pneumonia, and tuberculosis from chest X-ray images. The study showed that these models, especially EfficientNetB0, were accurate in rapid and fast classification, which could be helpful for radiologists in low-resource settings. However, the study brought out the limitations of dealing with large X-ray datasets and the constraints in the study with the dependency on pre-existing data [11]. Javid and Xie, 2024 "Lung disease classification using CNN-based deep learning models: A systematic review with focus on detection of small nodules for early diagnosis.". This work used data augmentation and transfer learning to maximize utility from a constrained dataset through accuracy maximization and diagnostic relevance. Class activation maps and segmentation methods supported this model, adding interpretability to these models towards clinical settings. Model transparency is indeed an asset in these environments. However, some of the issues that limit wider clinical use include a small dataset and high computational requirements, as given in Ref. [12].

Xiaoyi Liu et al. (2024) proposed a deep learning framework for lung disease classification from X-ray images using transfer learning and customized CNNs. MobileNetV2 was the baseline at 88.53 % accuracy, and fine-tuned MobileNet-FT reached 92.53%. A new model, MobileNet-Lung, made use of Squeeze-and-Excitation attention mechanisms to boost performance up to 93.33%. Although accuracy has thereby improved significantly, reliance on a single dataset and computational demands of the attention mechanism highlights the need for further validation for practical deployment [13].

III. ALGORITHMIC SURVEY

A. Deep Learning Models for Lung Disease Detection

Spatial hierarchies of images could easily be captured that turns out to be one of the cores for medical image analysis. In fact, different architectures of CNN are used, such as DenseNet, MobileNet, VGG-16, and InceptionV3. As the architecture of DenseNet suggests to reuse features from feed-forward connections between every layer and every other layer, DenseNet is particularly efficient for the identification of small patterns that have to be discovered in X-rays, like those determining bacteria from viral pneumonia. MobileNet: Light on resource and would classify images, detect the disease using fewer processing units; it also best suited to mobile and embedded applications. VGG-16 is computationally expensive and is often the perfect baseline in many investigations for the reason that it does an excellent job in extracting highly detailed features. InceptionV3 captures much more patterns within medical images at a reasonable trade-off concerning both accuracy and parameter efficiency.

Hybrids such as CNN with GRU or transformers improve capabilities towards capturing both spatial and temporal features. For example, the CNN-GRU model can detect progressive disease states by using CNN for spatial patterns and GRUs for temporal dependencies that would help distinguish between chronic and acute conditions. The latest research employs transformers, surpassing all previous architectures in modeling global dependencies. These enhance the performance of models by enabling them to better identify slight visual cues in the progress of a disease. For instance, CNN-GRU architecture uses GRU for learning temporal dependencies and CNN to learn spatial patterns that identify the progressive states of diseases. Some of the newer studies also introduce transformers, which are excellent in capturing global dependencies to better understand the subtle visual cues of lung illness over time.

The most widely known method is transfer learning, where a pre-trained model such as DenseNet201 is fine-tuned on specific medical datasets, like chest X-rays. This method works very well with foundational information from large datasets to speed up convergence and improve accuracy.

B. Data Augmentation Techniques

GANs have been used to create synthetic images for balancing class in medical datasets. Synthetic picture data creates through GANs improve the representation of rare scenarios and generalization for multi-label classification tasks. However, because poor quality images bring in noise degrading model accuracy, it is not very straightforward to produce high-quality synthetic images which pictorially could depict intricate lung features. It is very important to properly verify these artificial images. Through simulation of natural variation in medical images, additional augmentation techniques such as rotation, flipping, and contrast changes enhance the robustness of the model. Because lung diseases may manifest differently in different populations of patients and imaging scenarios, these techniques are particularly useful. Along with these methods, hybrid techniques that combine GAN-generated images with traditional augmentations are also emerging. Such strategies can produce more diversity and quality in the training data by utilizing the advantages of both approaches. The combination

TABLE I: COMPARATIVE ANALYSIS OF ALL STUDIED PAPERS

Sr.No	Title	Year	Methodology	Strengths and Weaknesses	
1	Multi-Label Classification of Lung Diseases Using Deep Learning	2024	Models: MobileNet, DenseNet, VGG-16. Dataset: NIH Chest X-ray 14. Data Augmentation: GANs	Strengths: GAN-based data augmentation, High classification accuracy. Weaknesses: Limited impact of GAN on performance, Synthetic data diversity issues	
2	A Survey of Deep Learning for Lung Disease Detection on Medical Images: State-of-the-Art, Taxonomy, Issues and Future Directions	2020	Analyzes 98 articles (2016-2020), categorized by image types, features, data augmentation, deep learning, transfer learning, ensemble classifiers, and lung disease types	Strengths: Comprehensive taxonomy of deep learning systems for lung disease detection. Weaknesses: Limited set of lung diseases	
3	An Explainable AI Model for Interpretable Lung Disease Classification	2021	Customized CNN with Layer-wise Relevance Propagation (LRP), Deep Taylor Decomposition (DTD), and LIME for explainability	Strengths: DenseNet201 accuracy: 84% Weaknesses: Model included irrelevant areas	Achieved explanations
4	Explainable Lung Disease Classification from Chest X-Ray Images Utilizing Deep Learning and XAI	2024	Models include CNN, hybrid models, ensembles. Methodologies: hyperparameter tuning, stratified k-fold cross-validation, transfer learning with fine-tuning	Strengths: Advanced models: CNNs, ensembles, transformers. High accuracy with Xception model. Weaknesses: Limited dataset diversity	
5	Collaboration between Explainable AI and Pulmonologists for Pulmonary Function Test Interpretation	2024	Two-phase design involving pulmonologists interpreting tests with and without explainable AI	Strengths: Significant improvement in diagnostic accuracy. Weaknesses: Small dataset	
6	Explainable AI in Deep Learning-based Medical Image Analysis	2024	Survey of XAI techniques in medical imaging, categorizing into visual, textual, and example-based explanations.	Strengths: Comprehensive overview of XAI techniques. Weaknesses: Need for non-visual explanations.	
7	Explainable AI Model for MultiLung Disease Classification Using Transfer Learning	2024	Fine-tuned DenseNet-201 with SHAP, LIME, Grad-CAM, GradCAM++ for explainability.	Strengths: Interpretability through SHAP, LIME, Grad-CAM. Weaknesses: Limited diversity in dataset sources.	
8	Enhancing Lung Abnormality Detection with CNN and GRU using XAI.	2024	DCNN for feature extraction + GRU for temporal dependencies. XAI with LIME, SHAP, Grad-CAM	Strengths: Provides interpretable insights. Weaknesses: Focus on two datasets, limiting generalizability	
9	Hybrid Deep Learning Model for Lung Disease Detection from X-Ray Images	2020	Combines CNN, VGG16, data augmentation, and spatial transformer network (STN) in VDSNet	Strengths: Improved accuracy with VGG16 and STN. Weaknesses: High training time, slow convergence.	
10	Detection and Classification Of Lung Diseases for Pneumonia and COVID-19	2023	Image preprocessing, quality enhancement, lung-specific ROI extraction, feature extraction, ANN, SVM, RNN-LSTM for classification	Strengths: High detection Accuracy, Robust preprocessing. Weaknesses: Computationally intensive	
11	Lung Disease Classification Using Deep Learning Models from Chest X-ray Images	2023	The study used CNN models (EfficientNetB0, Dense Net169, DenseNet201) with transfer learning and data augmentation on combined datasets to classify COVID-19, pneumonia, and tuberculosis from chest X-rays, evaluating performance with a confusion matrix.	Strengths: Achieved high accuracy and model efficiency, suitable for low-resource settings. Weaknesses: Data acquisition and preprocessing were costly and time-intensive; reliance on public datasets may affect model robustness across diverse clinical settings.	
12	Deep Learning for Classification Diseases.	2024	This study applies CNN-based deep learning, with data augmentation, segmentation, and transfer learning using VGG16 and ResNet50, to classify lung diseases. Model interpretability and accuracy were enhanced using class activation maps, with evaluation metrics including F1-score, accuracy, and ROC-AUC.	Strengths: High diagnostic accuracy for small nodules. Effective use of transfer learning for limited datasets. Improved interpretability for clinical use. Weaknesses: High computational demands may limit accessibility. Small dataset size could impact generalizability.	
13	Deep Learning for Lung Disease Classification Using Transfer Learning and a Customized CNN Architecture with Attention	2024	Tested 5 pre-trained models; fine-tuned MobileNetV2 (88.53%) with SE block to create MobileNet-Lung (93.33%).	Strengths: High accuracy, transfer learning, attention mechanism. Weaknesses: Limited dataset, high training time, no real-time efficiency analysis.	

strategy would better deal with class imbalances and variability in medical images, thereby improving model performance and generalization across different clinical scenarios.

C. Explainable AI (XAI) Techniques for Model Interpretability

Grad-CAM stands for Gradient-weighted Class Activation Mapping. Grad-CAM is one of the most popular approaches in visualizing parts of an X-ray image that most profoundly affect the AI model's decisions it will go on to make. Therefore, for example, in the context of a diagnosis for pneumonia, its application has highlighted inflamed regions; hence, its focus would be in accord with clinical evaluation and much more interpretable. LIME could be understood to explain why the model assigned that particular classification to an image but changes the input data to understand the importance of features. SHAP provides a more intuitive explanation with a clearer understanding about how each feature is contributing towards prediction. These methods, like Grad-CAM, LIME, and SHAP, together create layered explanations for clinicians, which they can cross-validate with greater confidence. The reason is that these technologies, via visual heatmaps from Grad-CAM, perturbation-based analysis through LIME, and feature contribution quantification by SHAP, will enable clinicians to understand not just what the model is predicting but also why it is doing so. These methods in tandem together will provide an all-round look at the decision-making process of the model and hence better collaboration between AI systems and healthcare professionals while making more informed clinical decisions.

D. Evaluation Metrics for Performance Measurement

Evaluation metrics are crucial for assessment of performance because of a model. Accuracy refers to correctness in overall prediction and is fundamental for clinical viability. For all those cases wherein false positives or false negatives matter, such as misclassified healthy lungs or failure to detect disease, precision and memory become the issues of utmost importance. Particularly useful in the case of unbalanced datasets, a balanced metric would be F1 Score, which combines precision with recall. One other aspect in testing the model's discrimination capability is through the Receiver Operating Characteristic's Area Under the Curve, particularly useful in multi-class disease diagnosis.

IV. GENERAL FRAMEWORK

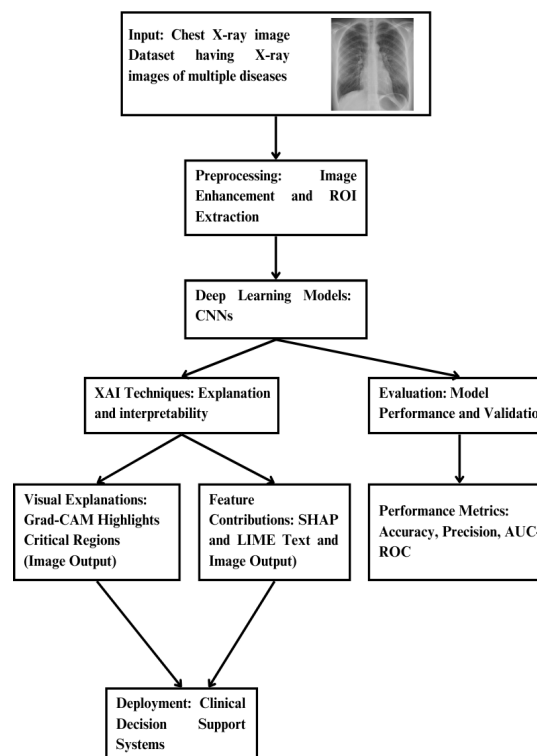


Figure 1: General framework for lung disease detection and classification.

This framework describes a holistic approach for the diagnosis of lung disorders from chest X-ray images using deep learning and explainable AI approaches. A Chest X-ray Dataset forms the basis for both model training and evaluation at the beginning of the process. Images with annotations for various lung disorders are included. Noise reduction and contrast modification methods of image enhancement are used for enhancing the quality of the images at the preprocessing stage. Moreover, Region of Interest Extraction ensures that the model focuses more on the lung regions, thereby enhancing its ability to detect the features peculiar to a disease. The core of the proposed framework is a Convolutional Neural Network, which is a deep learning architecture particularly designed for multi-class classification. By using its ability to recognize complex patterns in X-ray images, CNN is trained to detect and differentiate various lung conditions. For interpretability, XAI techniques are integrated. Grad-CAM produces heat maps showing where the important regions of the images are for making the predictions by the model. Thus, clinicians can more clearly see the focus points of the model. Further, SHAP and LIME explain at the feature level since they compute the contribution of each input feature to the decisions made by the model. LIME explains specific predictions locally while SHAP offers an overview of the significance of the features. All criteria like Accuracy, Precision, and AUC-ROC prove to have strong performance. In summary, the model is implemented in a Clinical Decision Support System in order to aid clinicians in arriving at a diagnosis of lung disease and understand why a prediction has been made so that trust may be built up for decision-making at the clinical level.

V. CHALLENGES

Developing deep learning-based explainable AI models for lung disease detection is challenging. Datasets are mostly biased and imbalanced towards the most common diseases and it becomes difficult to predict the rarest of conditions. Diseases that such currently built models cover are limited in scope, which are usually concentrated on tuberculosis, COVID-19, and pneumonia and thus lack clinical relevance for more complex scenarios in real-world situations. Another issue is overfitting, which makes the model less applicable to novel situations when there is not enough high-quality annotated data. Lastly, deep learning models could be computationally expensive and less likely to be adopted in under-resourced health care institutions in rural or poor settings.

VI. CONCLUSION

Integrate explainable AI techniques SHAP, LIME, Grad-CAM in deep learning models. These show hope in improving diagnostic precision, efficiency, and transparency concerning lung disease diagnosis. They do enhance interpretability, leading to increased trust amongst the medical fraternity as well as laying a path toward responsible AI-based medical imaging diagnostics. This step is essential towards developing scalable solutions for accurate lung disease diagnostics.

REFERENCES

- [1] M. Irtaza, A. Ali, M. Gulzar and A. Wali, "Multi-Label Classification of Lung Diseases Using Deep Learning," in *IEEE Access*, vol. 12, pp. 124062-124080, 2024, doi: 10.1109/ACCESS.2024.3454537.
- [2] Kieu, S.T.H.; Bade, A.; Hijazi, M.H.A.; Kolivand, H.A Survey of Deep Learning for Lung Disease Detection on Medical Images: State-of-the-Art, Taxonomy, Issues and Future Directions. *J. Imaging* 2020, 6, 131. <https://doi.org/10.3390/jimaging6120131>
- [3] V. Pitroda, M. M. Fouda and Z. M. Fadlullah, "An Explainable AI Model for Interpretable Lung Disease Classification," 2021 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS), Bandung, Indonesia, 2021, pp. 98-103, doi: 10.1109/IoTaIS53735.2021.9628573.
- [4] arXiv:2404.11428 [eess.IV] (or arXiv:2404.11428v1 [eess.IV] for this version) <https://doi.org/10.48550/arXiv.2404.11428>
- [5] Nilakash Das, et al., "Collaboration between explainable artificial intelligence and pulmonologists improves the accuracy of pulmonary function test interpretation" *European Respiratory Journal* May 2023, 61 (5) 2201720; DOI: 10.1183/13993003.01720-2022
- [6] Bas H.M. van der Velden, Hugo J. Kuijf, Kenneth G.A. Gilhuijs, Max A. Viergever, Explainable artificial intelligence (XAI) in deep learning-based medical image analysis, *Medical Image Analysis*, Volume 79, 2022, 102470, ISSN 1361-8415, <https://doi.org/10.1016/j.media.2022.102470>.
- [7] Eram Mahamud, et al., "An explainable artificial intelligence model for multiple lung diseases classification from chest X-ray images using fine-tuned transfer learning," *Decision Analytics Journal*, Volume 12, 2024, 100499, ISSN 2772-6622, <https://doi.org/10.1016/j.dajour.2024.100499>.

- [8] Md Khairul Islam, et al., "Enhancing lung abnormalities detection and classification using a Deep Convolutional Neural Network and GRU with explainable AI," *Machine Learning with Applications*, Volume 14, 2023, 100492, ISSN 2666-8270, <https://doi.org/10.1016/j.mlwa.2023.100492>
- [9] Bharati S, Podder P, Mondal MRH. Hybrid deep learning for detecting lung diseases from X-ray images. *Inform Med Unlocked*. 2020;20:100391. doi: 10.1016/j.imu.2020.100391.
- [10] Katarzyna Borys, et al., "Explainable AI in medical imaging: An overview for clinical practitioners," *European Journal of Radiology*, Volume 162, 2023, 1078 ISSN 0720-048X, <https://doi.org/10.1016/j.ejrad.2023.110787>.
- [11] Goyal S, Singh R. Detection and classification of lung diseases for pneumonia and Covid-19 using machine and deep learning techniques. *J Ambient Intell Human Comput*. 2023 Aug 25:1-16. doi: 10.1007/s12652-023-04899-8.
- [12] S. Sultana, A. Pramanik and M. S. Rahman, "Lung Disease Classification Using Deep Learning Models from Chest X-ray Images," 2023 3rd International Conference on Intelligent Communication and Computational Techniques (ICCT), Jaipur, India, 2023, pp. 1-7, doi: 10.1109/ICCT56969.2023.10075968.
- [13] Javid, M., & Xie, H. (2024). Deep Learning Models for Classification of Lung Diseases. *European Journal of Theoretical and Applied Sciences*, 2(2), 858-868. [https://doi.org/10.59324/ejtas.2024.2\(2\).77](https://doi.org/10.59324/ejtas.2024.2(2).77).