Multimodal Deep Learning System for Detecting Respiratory Diseases using Lung Imaging and Respiratory Audio

1st Vaishnavi Eswar
Department of Computer Science and Business Systems
Thiagarajar College of Engineering
Madurai, Tamil Nadu, India
vaishnavieswar@student.tce.edu

2ndMazwin Asra A
Department of Computer Science and Business Systems
Thiagarajar College of Engineering
Madurai, Tamil Nadu, India
mazwin@student.tce.edu

3rdPriya Thiagarajan
Department of Computer Science and Business Systems
Thiagarajar College of Engineering
Madurai, Tamil Nadu, India
ptca@tce.edu

4thSowndharya K
Department of Computer Science and Business Systems
Thiagarajar College of Engineering
Madurai, Tamil Nadu, India
sowndharyak@student.tce.edu

Abstract **-** This project develops a multimodal deep learning system aimed at the early and accurate detection of respiratory diseases by integrating two critical diagnostic modalities: chest CT/CXR imaging and stethoscope-based lung and cough audio recordings. Instead of depending on a single source of evidence, the proposed model simultaneously learns structural lung abnormalities from medical images and acoustic patterns from cough and breath sounds. Mel-spectrograms are generated from the audio signals, while feature extraction is carried out using a ResNet50 backbone for images, a custom CNN for spectrograms, and a Transformer network for MFCC-based temporal patterns. These features are integrated to improve diagnostic performance and Explainable AI methods such as Grad CAM and audio attention maps helps in highlighting the image regions and sound segments that shows an impact in model's decisions and contribute to clinical trust.

Keywords - Multimodal AI, lung disease detection, CT/CXR images, lung sounds, mel-spectrograms, deep learning, neural networks, explainable AI, Grad-CAM, MFCC, attention maps, medical diagnostics

I. INTRODUCTION

Respiratory diseases such as Pneumonia, Tuberculosis still be a major global health challenges particularly in developing countries where diagnostic services are limited. This become the major reasons for a large number of hospitalizations and their accurate diagnosis becomes difficult since the early symptoms often comes along with common respiratory illness. Although chest X rays, CT scans are highly used for finding lung problems, their effectiveness is often constrained by limited image quality and requires expert analysis. These limitations frequently delay diagnosis leading to disease progression and higher healthcare challenges.

Recent technological advances in digital imaging has created new opportunities for automated lung disease screening. However, most existing AI based systems rely on chest X ray classification, cough sound analysis or respiratory sound analysis. Such systems struggle when individual data sources are noisy and incomplete. Lung diseases often show by changes in lung structure and breathing therefore depending on a single type of data reduces the system's ability of prediction. This shows that multimodal diagnostic methods are essentially needed that combine both visual lung imaging with respiratory audio to provide a more complete view of lung function, detect structural abnormalities as well as functional disruptions in airflow. At the same time, explainable artificial intelligence (XAI) has become highly essential to ensure transparency by identifying the specific regions that influence model predictions.

Motivated by these opportunities, This study provides a multimodal deep learning framework for early detection of lung diseases by integrating chest imaging data with respiratory audio. The proposed approach aims in improving diagnostic performance, provide a more accurate diagnostic system suitable for clinical environments and support clinicians in making informed decisions.

II. LITERATURE SURVEY

Recent work in respiratory disease detection has widely studied both audio based and image based approaches due to advancements in deep learning and diagnostic techniques. Several studies have used AI and ML for analyzing cough sounds and chest imaging. One study proposed an automated cough based diagnostic system which used denoising, segmentation and MLP based spectral mapping but focused only on childhood pneumonia and was limited by relying only on audio data [1]. Another research combined chest X rays, CT scans and cough scalogram images within a single CNN framework but the conversion

of audio into static images lead to the loss of important information and increased the system overhead [2]. A multimodal approach for lung cancer detection involving VGG16 based imaging along with GRU processed MFCC audio signals was done in another research, but it lacked a propoer integration and was limited by dataset imbalance [3]. Further another work used CNN, SVM and transfer learning models on cough audio for respiratory disease diagnosis but these systems only told about unimodal analysis and suffered from dataset inconsistencies [14]. A survey on multimodal lung disease detection told the impact of AI based multimodal integration techniques but also told about the absence of large multimodal datasets and limited real world validation [7].

Overall, existing approaches face several limitations including small and imbalanced datasets, lack of multimodal feature integration, loss of temporal audio patterns and inadequate validation. These gaps highlight the need for a more comprehensive diagnostic framework that integrates both cough audio and medical imaging which thus supports efficient feature extraction and combine reliable multimodal integration mechanisms to improve the early and accurate detection of respiratory diseases.

III. Proposed Methodology

The The proposed system implements a multimodal deep learning approach that integrates chest imaging and respiratory audio for the early detection of lung diseases as illustrated in the workflow diagram inFig 1.

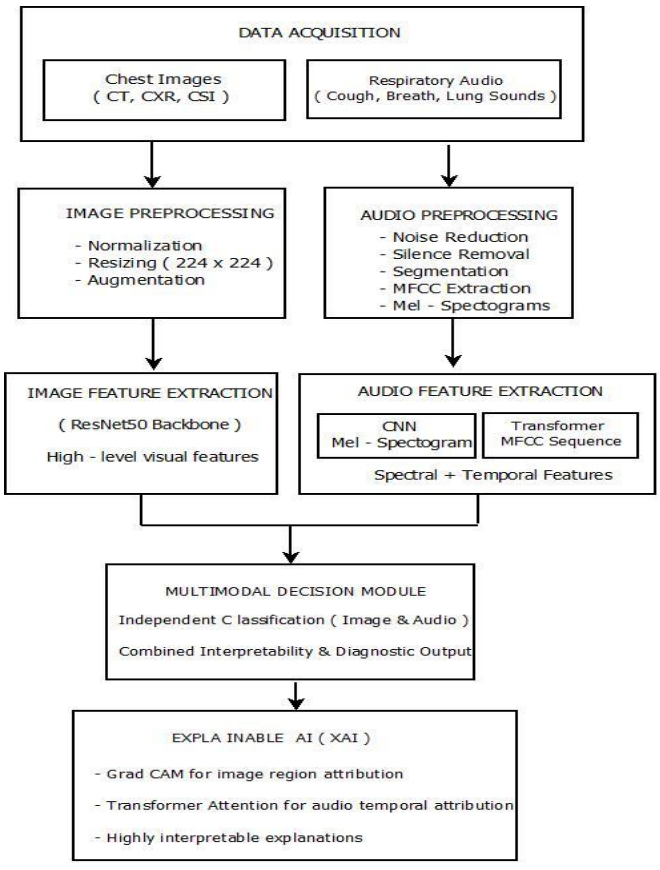

Fig. 1. Flow Diagram of the multimodal system

Two data sources which contained CT, CXR, CSI images and respiratory audio recordings were processed to capture both anatomical and physiological indicators of disease.

A. Audio Pipeline

Respiratory audio signals are preprocessed using noise reduction and silence removal to isolate meaningful cough or breath events. Two feature representations are derived: Mel-spectrograms for capturing spatial–frequency patterns and MFCC sequences for modeling temporal acoustic variationswith Mel-spectrograms computed using the Short-Time Fourier Transform$(STFT)$ as in eqn (1).

$$STFT(t,w) = \sum_{n=-\infty}^{\infty} x[n]\, w[n-t] e^{-jwn} \quad \text{.... eqn (1)}$$

A dual-branch architecture processes these features, where a CNN extracts spectral patterns from Mel-spectrograms, and a Transformer models long-range temporal dependencies in MFCC sequenceswith the CNN learning spectral structures through the convolution operationas in eqn (2) while the Transformer models long range temporal dependencies in MFCC sequences with MFCCs computed through the standard coefficient generation process as in eqn (3).

$$Y(i,j) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} X(i+m, j+n).\, K(m,n) \quad \text{.....eqn (2)}$$

$$MFCC(m) = \sum_{k=1}^{K} log\big(M(k)\big)\, cos\left[\frac{\pi m}{K}\left(k - \frac{1}{2}\right)\right] \text{....eqn (3)}$$

The outputs of both branches are integrated to form a comprehensive audio-based representation.

B. Image Pipeline

Chest images are resized, normalized and augmented to maintain uniformity and improve generalization. Visual feature extraction is performed using a pretrained ResNet50 model, selected for its strong residual learning capability and its imporved performance in medical imaging applicationswith feature refinement achieved through Residual Learning in ResNet50 as in eqn (4).

$$F(x) = \mathcal{H}(x) + x \quad \text{.... eqn (4)}$$

A lightweight classification head generates disease-specific predictions.

C. Classification and Decision Layer

The audio and image pipelines independently classify the given inputs into Normal, Pneumonia and Tuberculosis ensuring modularity, enabling use in multiple environments supporting audio-only by mobile auscultation, image-only or combined multimodal scenarioswith final class decisions optimized using the Cross-Entropy Loss used for classification as in eqn (5).

$$\mathcal{L} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c) \quad \text{.... eqn (5)}$$

This framework increases reliability when one modality is noisy, incomplete or unavailable.

D. Explainable AI

Explainable AI is integrated through two XAI techniques. Grad-CAM is applied to chest images to highlight the relevant lung regions such as opacities, cavitations and consolidation zoneswith the importance map computed using the Grad-CAM formulation, as in eqn (6). The correspondingweights used in eqn 6 are calculated as in eqn (7)

$$L^{Grad-CAM} = ReLU\left(\sum_k \alpha_k A^k\right) \quad \text{.... eqn (6)}$$

where

$$\alpha_k = \frac{1}{Z}\sum_i \sum_j \frac{\partial y}{\partial A_{ij}^k} \quad \text{.... eqn (7)}$$

For audio analysis, time segments such as high-intensity cough bursts or abnormal breath cycles are identified by Transformer attention weightswith the scaled dotproduct attention computed as in eqn (8).

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad \text{.... eqn (8)}$$

These above output effectively contribute to reliable and explainable clinical decision making.

E. Model Selection Rationale

The models were chosen to maximize the capability of each method like ResNet50.This was selected for image analysis due to its ability to capture even little abnormalities while maintaining a controlled gradient flow. CNN was used for audio to extract spectral features in Mel spectrograms since convolutional filters effectively capture distortion patterns which are also involved in respiratory conditions. The Transformer model was selected to process MFCC sequences since the attention mechanism has an improved performance in temporal dependencies and identifying critical areas. The combination of these models shows that spatial and temporal characteristics of respiratory diseases are effectively learned which supporting a reliable diagnostic framework.
By combining the feature extraction, this methodology ensures accurate predictions with increased reliability and shows multimodal AI as a practical solution for early detection of respiratory diseases across various clinical scenarios.

IV. Results and Inference

A multimodal framework was developed using both chest images and respiratory audio inorder to dedect respiratory disease detection . The below results show the system's performance across preprocessing, model training and explainability also showing its effectiveness in capturing both anatomical and physiological disease patterns.

A. Preprocessing pipeline

A preprocessing pipeline was applied to both audio and image modalities to improve the data quality and ensure consistency. These steps were highly crucial for reducing noise and improving the reliability of feature extraction.

1) Audio Preprocessing

A preprocessing pipeline was implemented for high quality inputs. Background disturbances were reduced by doing noise reduction using spectral subtraction which isolated cough and breath segments. The resulting Mel spectrograms and MFCCs showed clear frequencies with sharper coughs which provided the CNN and Transformer branches to extract more audio features.

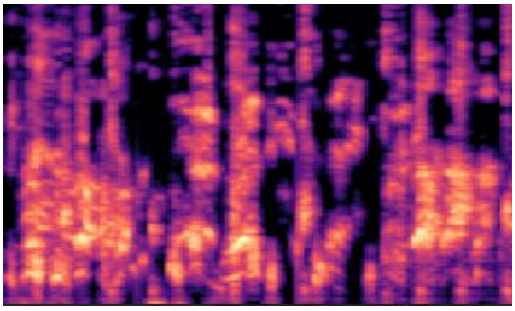

Fig. 2. Mel Spectogram for Tuberculosis Cough

2) Image Preprocessing

The CT, CXR and CSI images were resized to 224×224 pixels and normalized to ensure consistency. Variability was improved overfitting was reduced by performing data augumentation. The preprocessed dataset provided clarity making ResNet50 to better identify structural abnormalities such as opacities and cavitations.

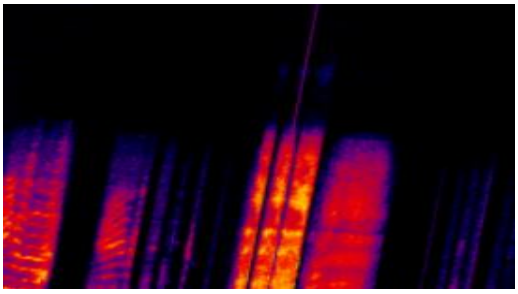

Fig. 3. CSI Image for Tuberculosis

Fig. 4. Chest X ray for Tuberculosis

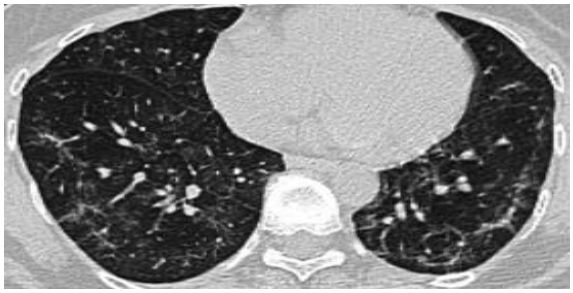

Fig. 5. CT Image for Tuberculosis

B. Audio Feature Learning and Model Evaluation

The audio architecture integrated CNN based Mel spectrogram learning with Transformer based temporal MFCC modeling was trained for 150 epochs. CNN captured spectral signatures such as cough distribution and the Transformer learned temporal patterns such as breath abnormalities, cough patterns. The integrated model achieved an accuracy of 88.37% predicting Tuberculosis, Pneumonia and Normal audio recordings. This indicated that the combination of spectral and temporal learning outperformed unimodal approaches.

Fig. 6. Audio Prediction for Tuberculosis

C. Image Feature Learning and Model Evaluation

The image classification model built using a ResNet50 backbone captured medical imaging signs across respiratory images. After training 150 epochs, the model achieved an accuracy of 93.57% which correctly classified Tuberculosis, Pneumonia and Normal conditions across test images confiming that the combination of preprocessing, standardized format recognizes early stage abnormalities from imaging modalities.

Fig. 7. Image Prediction for Tuberculosis

D. Explainable AI Evaluation

Explainable AI outputs were generated for both modalities to assess interpretability and model reliability which provided clear insight into the regions which helps understand how specific lung regions and acoustic events influenced the final diagnosis.

1) Grad CAM for Chest Images

Grad CAM visualization provided a heatmap which highlights the regions contributing to the decision making. For example in the Pneumonia cases, activation map highly minded the mid upper lung zones where patchy opacities commonly occur showing early signs of infection. This overlay confirms that the model focused on medically relevant structures.

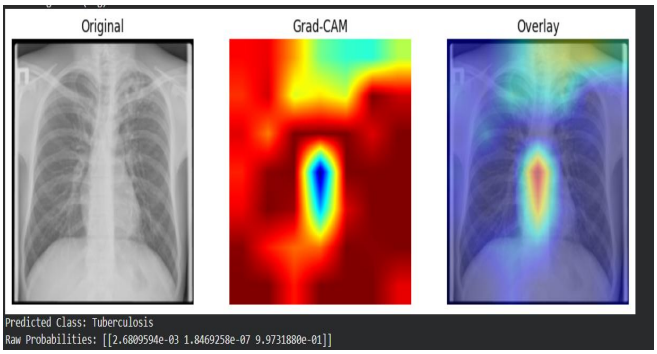

Fig. 8. Grad CAM Visualization for Tuberculosis

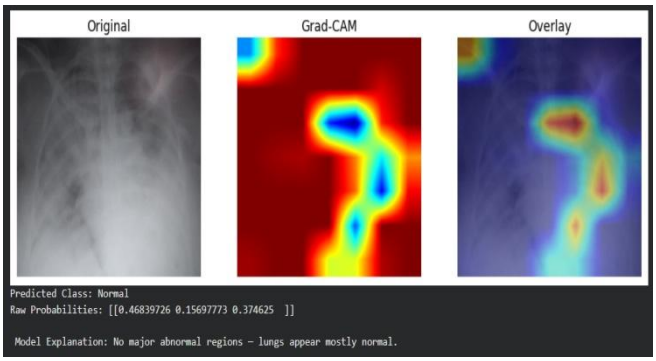

Fig. 9. Grad CAM Visualization for Normal

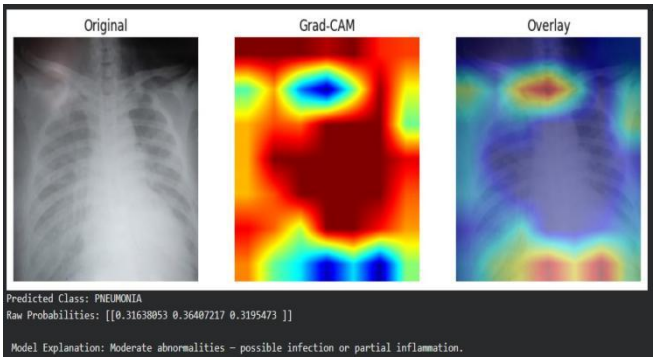

Fig. 10. Grad CAM Visualization for Pneumonia

2) Transformer Attention for Audio Signals

The Transformer attention matrix identified 449 sound events showing early sharp peaks. These segments refer to sudden cough bursts which contributed int the model's final predictions. These patterns are important as they can separate classify cough sounds as pathological or normal because of early and forceful cough bursts.

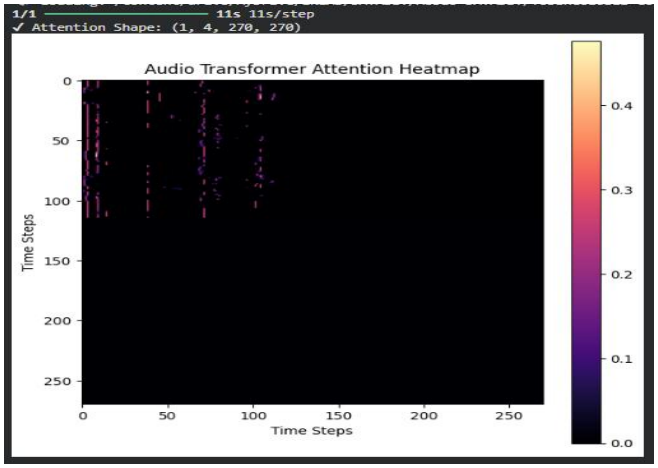

Fig. 11. Tuberculosis Audio Attention Heatmap

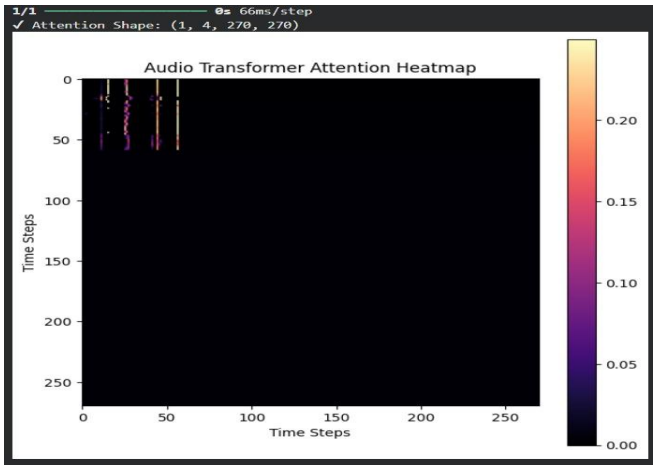

Fig. 12. Pneumonia Audio Attention Heatmap

E. Inference

Training was extended to 150 epochs since both models showed improvement in accuracy during the early and mid training stages which showed that the networks were still effectively learning audio and visual features. Towards the later epochs, the accuracy stabilized with further little improvement and the validation loss also stabilized indicating that 150 epochs provided the optimal balance between learning and preventing overfitting. At this stage, the models achieved their highest stable test accuracies which were 88.37% for audio and 93.57% for images.

These results confirm that the multimodal framework integrates respiratory audio and chest imaging for detection of Pneumonia, Tuberculosis and Normal conditions. The preprocessing improved data quality while the CNN Transformer audio model and ResNet50 image model gve an improved accuracy and captured spectral, temporal and structural disease patterns while Explainable AI validated the system by highlighting the meaningful lung regions and high energy cough events. Overall, the combined results show that the system is reliable, interpretable and meaningful.

V. Conclusion

The project successfully developed a multimodal deep learning framework for the early detection of respiratory diseases by integrating chest imaging and respiratory audio. The system system ensured consistent and high-quality inputs by including noise reduction, silence removal, Mel spectrogram, MFCC extraction and image normalization with augmentation. Feature extraction was performed using a audio architecture which integrated CNN and Transformer networks while features were derived using a ResNet50 architecture which was capable in capturing lung abnormalities across CT, CXR and CSI scans. The modular system allows diagnosis using audio, images even both making it reliable even under noisy conditions. Explainable AI methods such as Grad CAM and Transformer attention highlighted relevant lung regions and audio signals which enhanced reliability also enabling the system to learn both structural and sound features of Pneumonia, Tuberculosis and healthy lung conditions. Finally the framework shows the effectiveness of multimodal AI to support early and reliable respiratory disease detection and also addresses the limitations of unimodal approaches. Future work will focus on increasing the dataset size and applying the system to dedect more respiratory diseases which improve clinical use.

References

[1] Benaliouche, H., Hafi, H., Bendjenna, H., & Alshaikh, Z. (2025). Towards AI-Driven Cough Sound Analysis for Respiratory Disease Diagnosis. *IEEE Access*.

[2] Medeuova, Z. (2025). A Survey on Multimodal Approaches for Lung Disease Diagnosis using Deep Learning. *Journal of Emerging Technologies and Computing*, *1*(1).

[3] Ghori, K. W. U. R. (2025). *Multimodal Deep Learning for Lungs Cancer Detection: Integrating Audio and Image Analysis with Web-Based Accessibility* (Doctoral dissertation, Dublin, National College of Ireland).

[4] Zhang, P. F., Ting, H. N., & Chang, S. W. (2025). Using Cough Sounds for the Recognition of Multiclass Respiratory Diseases with Artificial Intelligence: A Review. *IEEE Access*.

[5] Abdullah, Fatima, Z., Abdullah, J., Rodríguez, J. L. O., & Sidorov, G. (2025). A multimodal AI framework for automated multiclass lung disease diagnosis from respiratory sounds with simulated biomarker fusion and

personalized medication recommendation. *International Journal of Molecular Sciences, 26(15), 7135.*

[6] Shokouhmand, S., Bhatt, S., & Faezipour, M. (2025). Artificial Intelligence in Respiratory Health: A Review of AI-Driven Analysis of Oral and Nasal Breathing Sounds for Pulmonary Assessment. *Electronics*, *14*(10), 1994.

[7] Malik, H., & Anees, T. (2024). Multi-modal deep learning methods for classification of chest diseases using different medical imaging and cough sounds. *Plos one*, *19*(3), e0296352.

[8] Kapetanidis, P., Kalioras, F., Tsakonas, C., Tzamalis, P., Kontogiannis, G., Karamanidou, T., ... & Nikoletseas, S. (2024). Respiratory diseases diagnosis using audio analysis and artificial intelligence: a systematic review. *Sensors*, *24*(4), 1173.

[9] Shehab, S. A., Mohammed, K. K., Darwish, A., & Hassanien, A. E. (2024). Deep learning and featurefusion-based lung sound recognition model to diagnoserespiratory diseases. *Soft Computing, 28(19), 11667–11683.*

[10] Alghamdi, N. S., Zakariah, M., & Karamti, H. (2024). A deep CNN-based acoustic model for the identification of lung diseases utilizing extracted MFCC features from respiratory sounds. Multimedia Tools and Applications, 83(35), 82871-82903.

[11] Kumar, S., Chaube, M. K., Alsamhi, S. H., Gupta, S. K., Guizani, M., Gravina, R., & Fortino, G. (2022). A novel multimodal fusion framework for early diagnosis and accurate classification of COVID-19 patients using X-ray images and speech signal processing techniques. *Computer methods and programs in biomedicine*, *226*, 107109..

[12] Aljaddouh, B., Malathi, D., & Alaswad, F. (2024). Multimodal Disease Detection and Classification Using Breath Sounds and Vision Transformer for Improved Diagnosis. *Procedia Computer Science*, *235*, 1436-1444.

[13] Sahu, P., Kumar, S., & Behera, A. K. (2024, July). SOUNDNet: Leveraging Deep Learning for the Severity Classification of Chronic Obstructive Pulmonary Disease Based on Lung Sound Analysis. In *2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)* (pp. 1-6). IEEE.

[14] Sharan, R. V., Qian, K., & Yamamoto, Y. (2023). Automated cough sound analysis for detectingchildhood pneumonia. *IEEE Journal of Biomedical and Health Informatics, 28(1), 193–203.*

[15] Kumar, S., et al. (2022). A novel multimodal fusion framework for early diagnosis and accurate classification of COVID-19 patients using X-ray images and speech signal processing techniques. *Computer Methods and Programs in Biomedicine, 226, 107109.*

[16] Lalouani, W., et al. (2022). Enabling effective breathing sound analysis for automated diagnosis of lung diseases. *Smart Health, 26, 100329.*

[17] Brunese, L., et al. (2022). A neural network-based method for respiratory sound analysis and lung disease detection. *Applied Sciences, 12(8), 3877.*

[18] Bhowmik, R. T. (2021). A multi-modal respiratory disease exacerbation prediction technique based on a spatio-temporal machine learning architecture. arXiv preprint arXiv:2103.03086.

[19] Jayalakshmy, S., & Sudha, G. F. (2021). GTCC-based BiLSTM deep-learning framework for respiratory sound classification using empirical mode decomposition. Neural Computing and Applications, 33(24), 17029-17040.

[20] Ott, J., Bruyette, D., Arbuckle, C., Balsz, D., Hecht, S., Shubitz, L., & Baldi, P. (2021). Detecting pulmonary Coccidioidomycosis with deep convolutional neural networks. Machine Learning with Applications, 5, 100040.