Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Explainable AI for Respiratory Disease Detection: Leveraging Deep Learning on Patient Audio Data

SRUJANA MADIRAJU (Fellow, IEEE), MANJULA K SHENOY(Senior Member,IEEE), DHANYA(Member, IEEE)

[1]Department of ICT, Manipal Institute of Technology, Manipal, Manipal Academy of Higher Education, Manipal, India
[2]Department of ICT, Manipal Institute of Technology, Manipal, Manipal Academy of Higher Education, Manipal, India (email: manju.shenoy@manipal.edu)
[3]Department of Occupational Therapy, Manipal College of Health Professions, Manipal Academy of Higher Education, Manipal, India

ABSTRACT Respiratory diseases affect millions of people around the world, making it necessary for reliable and interpretable diagnosis. Lung sound analysis is a non-invasive and cost-effective approach for detecting respiratory abnormalities such as wheezes and crackles, which are critical indicators of respiratory conditions like Chronic Obstructive Pulmonary Disease (COPD). This study uses machine learning techniques to detect crackles and wheezes from lung sounds automatically. Leveraging the respiratory sound Database, 13 Mel-Frequency Cepstaral Coefficients (MFCCs) were extracted from the audio recordings to classify respiratory abnormalities. While deep learning models achieve high accuracy, their black-box nature limits transparency. This study proposes an explainable AI (XAI) solution for respiratory disease classification using audio signals. This study ensures interpretability by identifying critical features influencing predictions by training models on publicly available datasets and incorporating Local Interpretable Model-Agnostic Explanations (LIME). Explainability analysis revealed critical features influencing predictions, ensuring model transparency. This research advances the development of trustworthy AI-driven diagnostic tools, contributing to enhanced transparency in healthcare.

INDEX TERMS Audio classification, COPD, Deep Learning, Explainable AI (XAI), Machine Learning, LIME, Healthcare Diagnostics, Mel-Frequency Cepstral Coefficients (MFCCs), Crackles, Wheezes

I. INTRODUCTION

Millions of people die worldwide due to Chronic Obstructive Pulmonary Disease (COPD). This stands fourth among the leading causes of death worldwide killing 3.5 million in 2021, which is approximately 5% of global deaths [1]. 90% of deaths are under the age of 70 and from low and middle-income families of countries (LMIC). COPD is the eighth leading cause of poor health worldwide (measured by disability-adjusted life years). About 70% of COPD cases in high-income [1] countries occur due to tobacco smoking. In LMIC, smoking leads to 30 to 40% of COPD cases, and air pollution is a major risk factor. COPD is sometimes called emphysema or chronic bronchitis causing restricted airflow into the lungs and breathing problems. The lungs get damaged or clogged with phlegm causing symptoms such as cough, phlegm, breathing difficulty, wheezing, and tiredness. People with COPD are at a higher risk of other health problems. It is not curable, but it can be improved by avoiding smoking and exposure to pollution. Affected people can be vaccinated to prevent infections. Treatment exists with medicines, oxygen, and pulmonary rehabilitation. By leveraging deep learning, this paper outlines a method for detecting sounds that indicate respiratory diseases such as COPD in a non -invasive way by analyzing recorded audio of respiratory noises and makes it usable and interpretable to the user with the help of XAI. Section II analyzes the dataset, III related work, IV gives methodology details, V gives XAI integration with LIME, VI conclusion followed by future work in the VIIth section.

A. ENHANCING INTERPRETABILITY, TRANSPARENCY AND EXPLAINABILITY

Explainable AI (XAI) is emerging as a critical enabler of trust, transparency and reliability in healthcare. Deep Learning models are considered black boxes that lack proper interpretability, making it challenging for professionals to trust

their applications. XAI bridges the gap by providing insights into how the models arrive at their decisions. Techniques such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) help uncover the influence of features on predictions, allowing clinicians to validate AI-driven outcomes against clinical knowledge. One prominent application of XAI in healthcare is in the diagnosis and treatment planning of diseases. Moreover it promotes fairness and reduces biases. AI models, when developed without proper transparency, may unwittingly learn biased patterns from the data, leading to inequitable care. Potential biases in decision making can be identified and addressed using explainability. Transparency, interpretability and explainability are closely related yet distinct concepts in the context of AI systems. These terms are often used interchangeably, leading to confusion, particularly in discussions about AI ethics, accountability, and trust. While they overlap, each addresses unique aspects of making AI systems understandable and trustworthy to humans. Transparency focuses on providing evidence and documentation regarding an AI systems's decisions and operations. It helps us validate and verify the outcomes. Examples of transparency tools include model cards, dataset documentation, fairness indicators, and algorithmic impact asssessments. Interpretability focuses on designing AI models that are inherently understandable. Unlike opaque models requiring additional methods for interpretation, interpretable models are self-explanatory, enabling users to comprehend their functionality directly. Examples include decision trees, linear models, and monotonic neural networks. Explainability extends beyond interpretability by aiming to make opaque models, such as deep learning networks, understandable after they are trained. Explainability involves creating post-hoc techniques to explain a model's outputs and behaviors. Common methods include SHAP and LIME as mentioned above.

B. LIME

LIME (Local Interpretable Model-agnostic Explanations) is a technique made to make AI models more transparent and interpretable. LIME addresses the challenge of machine learning models being black-boxes providing a way to make predictions clear and understandable. The core idea of LIME is to focus on a single prediction made by the model by creating a simpler model (like a decision tree or a linear model) that mimics the behavior of the original complex model. But it does this only for that particular prediction. This allows users to understand which features (inputs) were most contributing in the model's prediction process for a particular case. For instance, in medical diagnosis, if an AI model predicts that a patient is at high risk of having a respiratory disease, LIME can highlight which factors (like patient age, breathing patterns, exposures) influenced most to this prediction. One of LIME's key strengths is its model-agnostic nature. It can be applied to any machine learning model regardless of it's complexity. In summary LIME bridges the gap between complex AI and human understanding, enabling stakeholders to gain insights into how decisions are made. Section V explains more about LIME techniques integrated with the AI model.

C. ADDRESSING BIAS AND PROMOTING FAIRNESS

Beyond integrating trust, XAI facilitates more equitable and fair healthcare practices. AI models trained on biased data can disproportionately impact underrepresented groups. By introducing transparency in the decision-making process, XAI helps uncover these biases. For example, explainable models can ensure that diagnostic tools do not overlook critical symptoms in minority populations due to biased data patterns. An important concept closely related to explainability is AI alignment. It is the concept of aligning AI systems with human values, ethics and intentions which makes those systems more powerful and autonomous. Misaligned AI systems, even those with seemingly minor discrepancies in their objectives or training processes, can produce unintended or harmful outcomes.

D. FACILITATING PERSONALIZED MEDICINE AND COLLABORATIVE CARE

In addition to improving fairness, XAI plays a pivot role in advancing personalized medicine. By addressing how each characteristic influence predictions, XAI supports treatment plans that considers each patient's unique profile. This capability extends to areas such as drug discovery where explainable models can highlight specific biological features linked to treatment efficiency. The importance of XAI systems cannot be overstated, especially in high-stakes domains such as healthcare. Firstly, it builds trust and accountability. Secondly, it aids in identifying biases. Finally explainablility promotes scientific discovery. By clarifying model behaviour, researchers can gain insights into the topic being studied.

E. RESPONSIBLE AI

Responsible AI is a broader framework for developing trustworthy and accountable AI systems which is termed as AI alignment. Responsible AI in healthcare focuses on social values, including privacy, safety and well-being. For example, when an AI system recommends a diagnosis or treatment, explainability allows clinicians to trace it's reasoning, ensuring the decision is unbiased and robust. This level of transparency is important in domains like healthcare. This forsters more proximity to realizing the full potential of AI while safeguarding ethical principles and improving patient care.

II. ABOUT THE DATASET

Respiratory sounds serve as key indicators of lung health and are widely used in diagnosing respiratory conditions. The sounds produced during breathing are influenced by airflow patterns, lung tissue characteristics, and the presence of secretions. For instance, wheezing is often associated with obstructive airway diseases such as asthma and chronic obstructive pulmonary disease (COPD).

With advancements in digital stethoscopes and audio recording techniques, respiratory sounds can now be captured as digital data. This opens the door for machine learning applications, enabling automated diagnosis of respiratory disorders like asthma, pneumonia, and bronchiolitis.

The *Respiratory Sound Database* was developed through a collaboration between research teams in Portugal and Greece. It consists of 920 recordings of varying durations, ranging from 10 to 90 seconds, collected from 126 patients. The dataset includes approximately 5.5 hours of respiratory audio, covering 6,898 respiratory cycles—among which 1,864 exhibit crackles, 886 contain wheezes, and 506 feature both. The recordings include both clean and noisy samples, simulating real-world clinical conditions.

Additionally, the dataset provides useful metadata such as patient identifiers, recording locations (e.g., trachea, chest, back), and associated health conditions. This supplementary information is valuable for multi-class classification tasks and the development of personalized diagnostic models.

III. RELATED WORK

The majority of the literature is centered on images. Analysis of respiratory sounds is less traceable, but there are existing works [15] where a computer system was designed for the easy analysis of lung sounds using the package DasyLAB. [14] used Deep Learning architectures such as VGGish, YAMNet, and CNN-LSTM to automate the process. Further studies were conducted on spirometry data [18]. [3] developed a machine learning framework to predict COPD exacerbations using data from personal air quality monitors, health records, and lifestyle information. Explainable AI methods like SHAP were integrated to highlight pollutant exposure and clinical factors as key predictors, emphasizing the heterogeneity in exacerbation risks across patient groups. Studies on specific respiratory diseases were also conducted, such as Asthma [11] [12]. Explainable AI frameworks were employed to predict the risk factors of COPD in smokers [20]. One notable study [5] examined a cohort of 78 patients using an electronic nose (e-nose) system integrated with eight sensors to analyze exhaled air for detecting COPD. The methodology involved transfer learning techniques based on Deep Neural Networks (DNNs), using five pre-trained Convolutional Neural Networks (CNNs), including GoogleNet, which achieved a test accuracy of 100%. The system used a twin Case-Based Reasoning (CBR) model to interpret DNN results. The method converted the black-box model into a more interpretable white-box framework. The e-nose sensor array was considered as the input for supervised classification algorithms, which classified exhaled breath samples from healthy controls, COPD patients, and smokers by analyzing volatile organic compounds (VOC). This study highlights the potential of sensor-based data and transfer learning approaches for COPD detection, distinguishing itself from our work, which focuses on analyzing respiratory sounds using audio datasets and XAI techniques. Another study integrated XAI with its deep learning models on chest X-ray and CT scan datasets to diagnose airway diseases like tuberculosis, pulmonary embolism, and pneumoconiosis [9]. XAI in this study was integrated through a graphical interface that provides interpretable predictions, along with treatment and preventive insights that improve user trust. The best performing model that achieved a high accuracy of 99.15% in disease detection was InceptionResNetV2. In conclusion these studies mentioned in this section demonstrate the growing integration of deep learning and explainable AI in respiratory disease detection. While prior research has explored image-based diagnostics and sensor-driven approaches, our work uniquely focuses on analyzing respiratory sounds using deep learning and XAI techniques. By leveraging audio data, we aim to provide a non-invasive, cost-effective alternative for respiratory disease diagnosis, contributing to the broader landscape of AI-driven healthcare solutions.

IV. METHODOLOGY

A. DATA COLLECTION

The data used in this paper is obtained from the ICBHI Respiratory Sound Database hosted on the Kaggle website. This dataset comprises of respiratory sound recordings collected from clinical environments, having a diverse range of patients with varying respiratory sounds. The recordings were acquired using stethoscopes positioned in different locations on the chest wall, ensuring a broad spectrum of sound data. Each audio file in the dataset is systematically named using five distinct elements, separated by an underscore (_), providing detailed metadata about the recording.

1) **Patient Number:** A unique identifier assigned to each patient the range being `101` to `226`.
2) **Recording Index:** The sequential number representing the specific recording of the patient.
3) **Chest Location:** The location on the chest where the recording was captured. The possible values are:
   - Trachea (`Tc`)
   - Anterior left (`Al`)
   - Anterior right (`Ar`)
   - Posterior left (`Pl`)
   - Posterior right (`Pr`)
   - Lateral left (`Ll`)
   - Lateral right (`Lr`)
4) **Acquisition Mode:** Specifies whether the recording was obtained using:
   - Sequential/single-channel mode (`sc`)
   - Simultaneous/multi-channel mode (`mc`)
5) **Recording Equipment:** Indicates the type of stethoscope or microphone used during the recording. The options are:
   - AKG C417L Microphone (`AKGC417L`)
   - 3M Littmann Classic II SE Stethoscope (`LittC2SE`)
   - 3M Littmann 3200 Electronic Stethoscope (`Litt3200`)
   - WelchAllyn Meditron Master Elite Electronic Stethoscope (`Meditron`)

Additionally, corresponding annotation files are provided for each audio recording. These annotation files contain four columns of information:

- **Beginning of the Respiratory Cycle (s):** The timestamp indicating the start of the respiratory cycle.
- **End of the Respiratory Cycle (s):** The timestamp indicating the end of the respiratory cycle.
- **Presence/Absence of Crackles:** Binary value denoting the presence (`1`) or absence (`0`) of crackles in the recording.
- **Presence/Absence of Wheezes:** Binary value denoting the presence (`1`) or absence (`0`) of wheezes in the recording.

This dataset structure provides a comprehensive framework for both training and evaluating machine learning models aimed at respiratory disease detection.

TABLE 1. Audio File Naming Explanation

Audio File Name	Description
`101_1b1_Al_sc_Litt3200.wav`	Patient 101, recording 1, anterior left, single-channel, Littmann 3200 stethoscope
`102_1b1_Pl_mc_Meditron.wav`	Patient 102, recording 1, posterior left, multi-channel, Meditron stethoscope
`103_1b1_Tc_sc_AKGC417L.wav`	Patient 103, recording 1, trachea, single-channel, AKG C417L microphone
`104_2b2_Lr_sc_LittC2SE.wav`	Patient 104, recording 2, lateral right, single-channel, Littmann Classic II SE

TABLE 2. Annotation File

Start Time (s)	End Time (s)	Crackles (0/1)	Wheezes (0/1)
0.00	0.95	1	0
1.02	2.10	0	1
2.15	3.25	1	1
3.30	4.00	0	0

B. DATA PREPROCESSING AND FEATURE ENGINEERING

In this step, we combine the audio files with their corresponding annotation files to generate a dataset containing relevant features, labels, and time intervals. The preprocessing workflow is as follows:

1) **Audio and Annotation File Pairing:** For each annotation file, the corresponding audio file is located by matching filenames, where the annotation file is of the format:

   `<patient_id>_<recording_index>_<chest_location>_<acquisition_mode>_<equipment>.txt`

   and the audio file is in WAV format.
2) **Audio Segmentation:** We extract the segments of audio corresponding to the start and end times specified in the annotation files. These segments are used for feature extraction.
3) **Feature Extraction:** For each audio segment, Mel-frequency cepstral coefficients (MFCCs) are computed using the `librosa` library. We take the mean of the MFCC coefficients across time frames to reduce dimensionality. This results in a feature vector for each audio segment.
4) **Label Assignment:** The labels for each segment are determined based on the presence of crackles and wheezes in the annotations. If both crackles and wheezes are present, the label is *both*; if only crackles are present, the label is *crackle*; if only wheezes are present, the label is *wheeze*; otherwise, the label is *neither*.
5) **Dataset Construction:** The dataset is constructed by storing the extracted features, labels, and time intervals in a data frame, which includes columns for patient ID, start time, end time, extracted features, and label.

The final dataset is saved as both a CSV file and a pickle file for ease of further processing and analysis.

TABLE 3. Preprocessed Dataset Summary

Patient ID	Start Time (s)	End Time (s)	Features	Label
101	0.00	1.23	[0.56, 1.23, ..., 0.89]	Crackle
102	2.34	3.56	[0.78, 1.45, ..., 1.02]	Wheeze
103	0.00	1.50	[1.02, 2.45, ..., 1.56]	Both
104	4.00	5.67	[0.67, 0.89, ..., 1.12]	Neither

C. MODEL TRAINING

The dataset was imbalanced with the following value counts for each class: *neither* (3,642 samples), *crackle* (1,864 samples), *wheeze* (886 samples), and *both* (506 samples). Due to this imbalance, the model may be biased toward the dominant class (*neither*). Initially, we trained the model without employing any class-balancing techniques to evaluate its performance under raw data conditions. Label Encoder was used to encode the categorical data into lables to facilitate classification. A StandardScaler was applied to normalize the feature values for ensuring uniformity across input features, which is essential for machine learning models.

1) Training the Convolutional Neural Network

A Convolutional Neural Network (CNN) was employed for the classification task composing of the following architecture:

1) **Input Layer:** The features extracted from MFCCs, were fed into a 1D convolutional network. The input

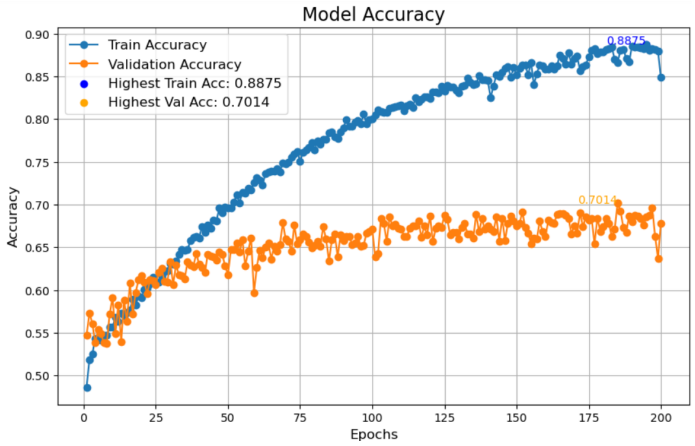


FIGURE 1. Training and Validation accuracy before SMOTE

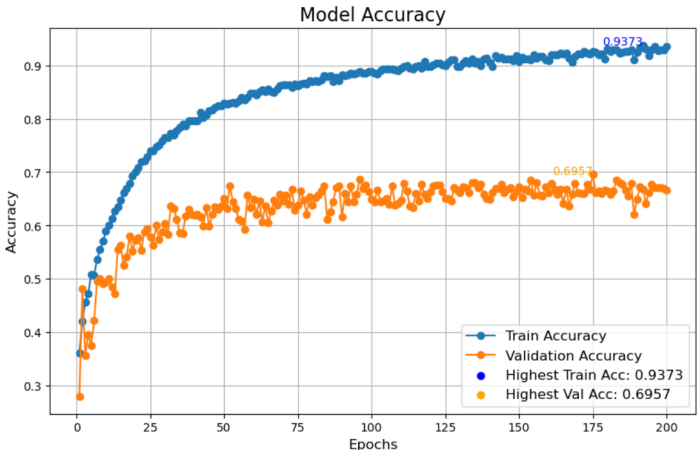


FIGURE 2. Training and Validation accuracy after SMOTE

shape for the model was `(number of features, 1)`, where the feature dimension varied depending on the preprocessing stage.

2) **Convolutional Layers:**
   - The first layer consists of 64 filters with a kernel size of 3, followed by a ReLU activation function.
   - A MaxPooling layer with a pool size of 2 used for down-sampling.
   - the second layer constitutes of 128 filters with a kernel size of 3, again followed by a ReLU funtion.

3) **Fully Connected Layers:**
   - The feature maps were passed through a dense layer with 128 neurons and a ReLU activation.
   - The final layer consists of 4 neurons, corresponding to the 4 classes to be implemented in a softmax activation fucntion for multi-class classification.

The model was compiled using the Adam optimizer. The dataset was divided into 80% as training set and the rest 20% as the validation set.

- A batch size of 32 used for manageable memory usage during training.
- 200 epochs, allowing the model to learn from the data.
- Validation data used to understand the performance and reduce overfitting.

2) Initial Observations

The CNN model performed well on the majority class (*Neither*) but showed less performance with the minority class (*Crackle*), (*Wheeze*), (*Both*). Figure 1 shows the accuracy plot of the model before SMOTE was applied. The highest validation accuracy was about 70.14% and the highest training accuracy being 88.75%, which clearly makes it evident that the model is overfitting. As a result, the Synthetic Minority Oversampling Technique (SMOTE) in subsequent training phases to improve the model's ability to detect minority classes.

D. APPLYING SMOTE

After applying SMOTE, the validation accuracy reached 69.5%. While this is slightly less than the model trained without SMOTE, it demonstrates slightly better generalization despite having a marginally lower validation accuracy. It is evident in Figure 2 that the gap between training and validation curves is smaller. While the validation accuracy is slightly lower, the generalization seems better as the model is not heavily overfitted.

E. MACHINE LEARNING MODELS

Traditional machine learning models were evaluated on this dataset which are:

- Random Forest (RF)
- Support Vector Machine (SVM)
- k-Nearest Neighbors (kNN)

1) Observation

Table 4 shows the accuracy of the models mentioned above. The SVM model performed poorly compared to the other models. The model was able to perform well with the *Wheeze* class but not with the other classes. This is clearly understood by the confusion matrix shown in Figure 3. In contrast, the Random Forest and the k-Nearest Neighbors demonstrated better balance with the Random Forest, achieving more accurate predictions for the *Wheeze* and *Crackle* classes. Figure 4 and Figure 5 show the confusion matrices of Random Forest and k-Nearest Neighbors models.

TABLE 4. Machine Learning Model Results

Model	Accuracy
Random Forest	73%
Support Vector Machine	53%
k-Nearest Neighbors	73%

Figure 6 shows the Reciever Operating Characteristic (ROC) and the Area Under the Curve (AUC) curve comparing the three models used to train the dataset. The SVM

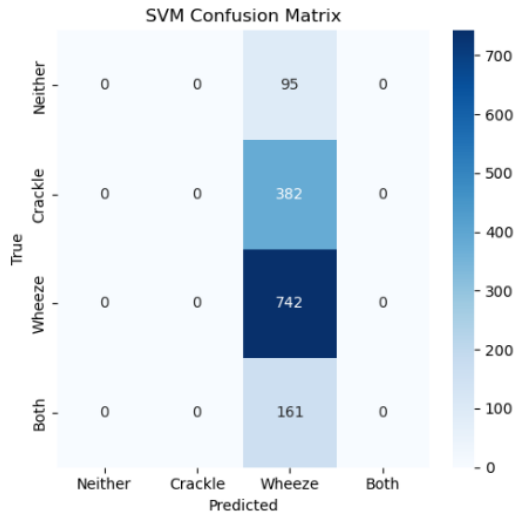


FIGURE 3. SVM confusion matrix

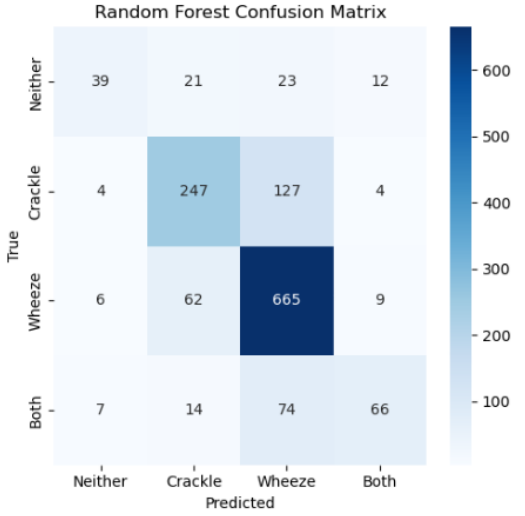


FIGURE 4. RF confusion matrix

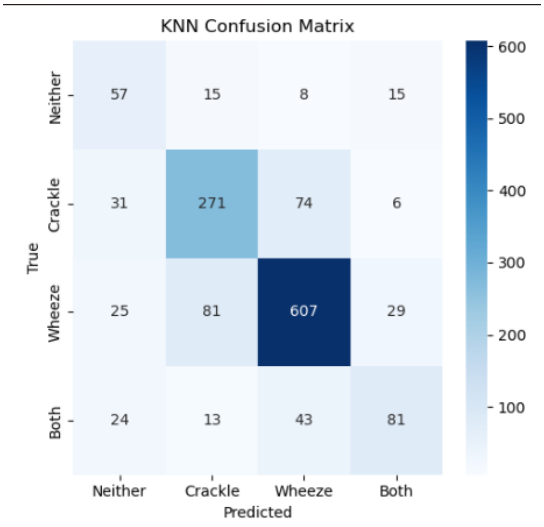


FIGURE 5. kNN confusion matrix

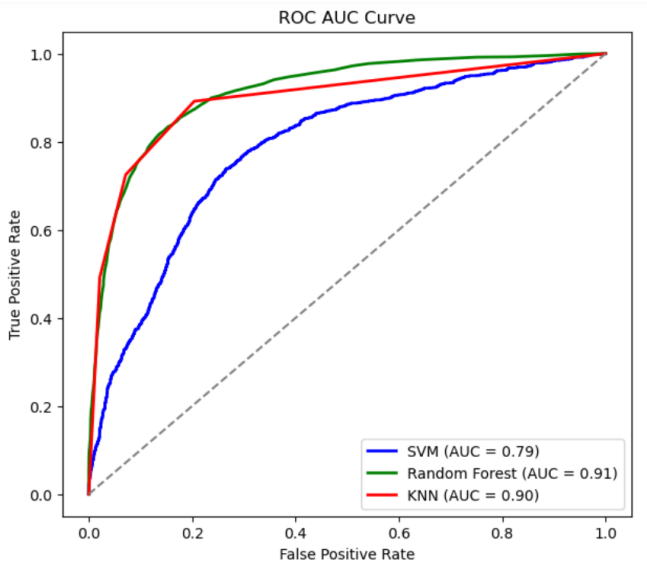


FIGURE 6. Comparison of Model performance by ROC AUC Curve

model with an AUC score of 0.79 indicates low performance. The Random Forest model achieved the highest AUC of 0.91 indicating good discriminative ability.

F. STACKING ENSEMBLE MODEL

To achieve a better classification result, a stacking ensemble model was trained as the final approach for the respiratory sound classification. The model used Support Vector Machine (SVM) and k-Nearest Neighbors (kNN) as base learners, with their predictions combined by a Logistic Regression meta-classifier. The SVM was configured with an RBF (Radial Basis Function) kernel and optimized hyperparameters $C = 10, \gamma = 1$, while kNN employed the Euclidean distance metric with k=5 and distance-based weighting. The data set was pre-processed, including feature scaling using StandardScalar and label encoding to ensure compatibility with the models. The stacking model achieved an accuracy of 78% outperfroming the individual base learners. The results summarized in the confusion matrix shown in Figure 7, highlighting the model's ability to balance predictions across classes effectively.

To further assess the performance of the stacking classifer, a multi-class ROC curve was plotted, and AUC scores were calculated for each class. The AUC values are shown in the Figure 8. These values indicate that the model was successful in differentiating between the four classes. It indicates a stronger ability to rank positive predictions higher than negative ones. The nearly perfect performance for Class 0 and strong scores for the other classes suggest that the model is highly effective at multi-class classification in the respective problem domain. Class 0, Class 1, Class 2, Class 3, Class 4 indicate *Neither*, *Crackle*, *Wheeze* and *Both* respectively.

V. INTEGRATING EXPLAINABLE AI

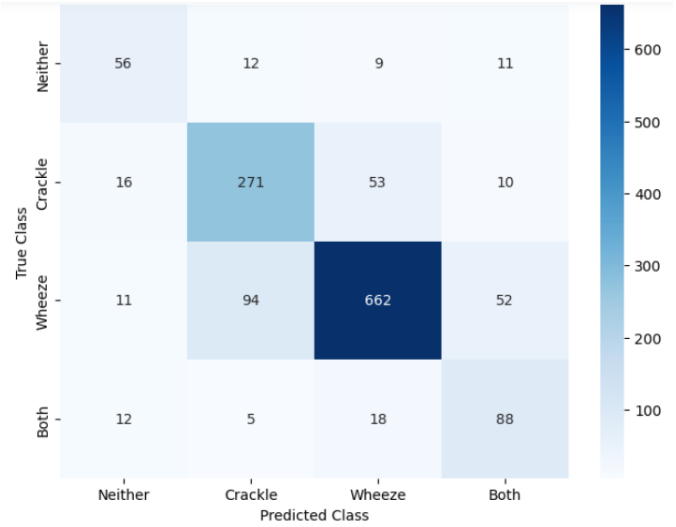


FIGURE 7. Stacking model confusion matrix

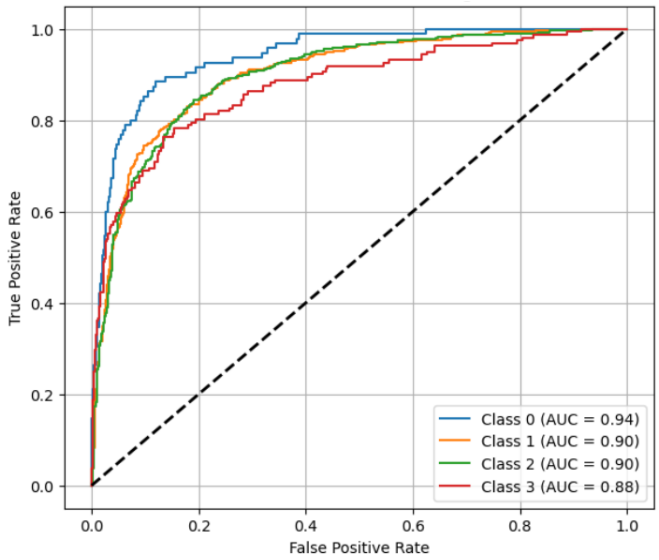


FIGURE 8. ROC AUC curve for Stacking model

A. PERMUTATION FEATURE IMPORTANCE

To improve the interpretability of the stacking model, the permutation feature importance was employed which is a widely used explainability technique. It specifies the contribution of each feature to the model's predictive performance. The values of individual features in the dataset are iteratively shuffled and the resulting decline in the model's performance is measured, estimating the relative importance of each feature. The results revealed significant variation in feature contributions. A bar plot illustrating the ranked feature importance (Figure 9) shows the prominent predictors, providing valuable insights into the model's prediction. The bar plot represents the permutation importance of the 13 MFCC features used in the stacking model. Each bar corresponds to a specific feature (indexed from 0-12), and the height of the stacking model's decision making process.

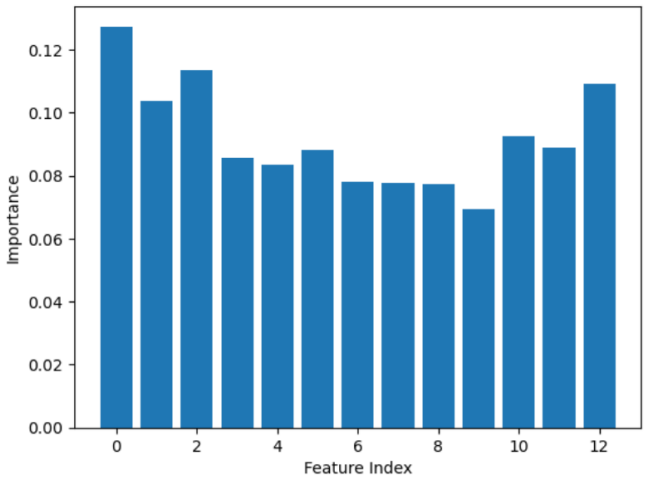


FIGURE 9. Permutation Feature Importance

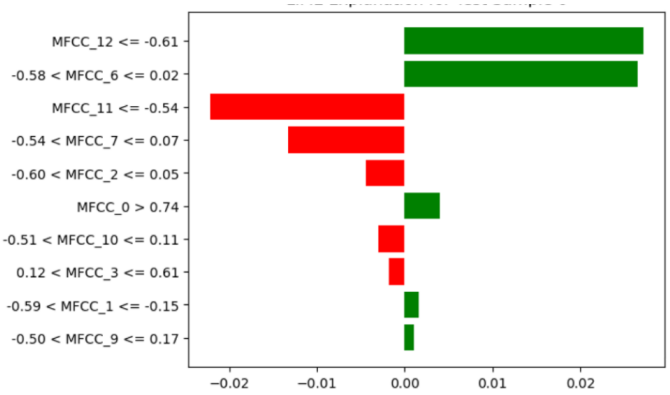


FIGURE 10. LIME Explanation for *crackle class*

Higher the bar, greater the impact of that particular feature on the model's performance.

B. LIME

The LIME visualization for one of the test samples belonging to the *crackle* class shows the contribution of MFCC features to the model's prediction. Green bars represent features that positively affected the predicted class, while red bars signify features that negatively influence the prediction. In the Figure 10 MFCC_12 and MFCC_6 strongly support the prediction, whereas MFCC_11 and MFCC_7 detracted from it. This is the local feature importance for a single test instance.

The second type of LIME visual representation is the LIME Dashboard. It is a more comprehensive breakdown of the LIME'S output showing the following:

1) Prediction probabilities of all the classes.
2) The most significant features for the predicted class and how they differentiate the predicted class from others.
3) The feature values with their ranges and weights.

This type of representation provides a global perspective on the predicted probabilities and their relative strengths. The

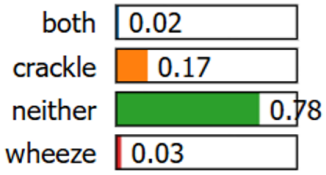


FIGURE 11. LIME predicted probabilities

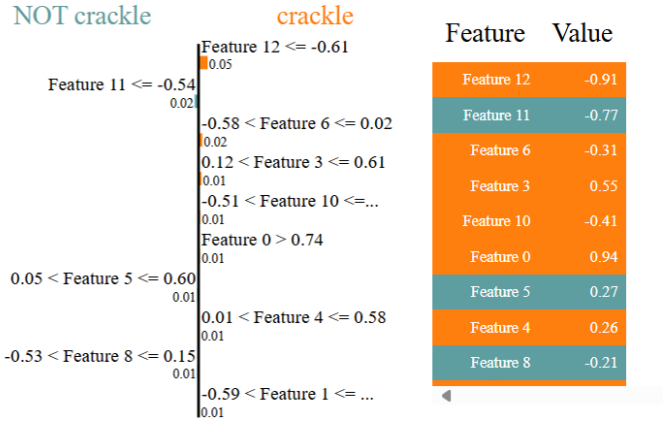


FIGURE 12. LIME Explanation for *Neither* class

bar chart in Figure 11 displays the prediction probabilities, where *Neither* has the highest probability of 0.78. The Figure 12 shows the feature contributions. The orange bars show the features responsible for prediction of *Crackle* class, while green bars indicate features that supported the 'Not Crackle' class (can be *Neither* or *Wheeze*). The right table gives the details of the MFCC feature valued for this sample. This visualization demonstrates how individual features influenced the model's prediction, enhancing the interpretability of the classification results. The dashboard clearly outlines the feature contributions for and against this prediction. The features are divided into two categories: those supporting the *Neither* class and those supporting other classes like *Crackle*. The key features supporting *Neither*:

1) Feature 11 exhibited a strong positive contribution towards this class.
2) Feature 5 and Feature 8 also provided moderate support due to respective ranges of values.

The key features opposing *Neither* (Supporting other classes) :

1) Feature 12 strongly favoured the *Crackle* class.
2) However, its contribution was not sufficient to override the combined effect of the dominant features supporting *Neither*.

VI. CONCLUSION

The integration of XAI methods significantly enhanced the interpretability and transparency of the respiratory sound classification model. The permutation feature importance highlighted the most influential features responsible for the model's prediction, providing a clear understanding of their contribution. LIME was further incorporated to improve the analysis by offering instance-specific insights, showing the role of individual features in understanding class probabilities. These techniques help in improving not only trust but also decision-making process. These methods provide actionable insights to refine the model and ensure robust performance especially in the domain of medicine. The approach was helpful to classify respiratory sounds into four categories *Crackle*, *Wheeze*, *Both* and *Neither* which are critical in understanding the nature of respiratory anomalies such as crackles, wheezes that often indicate conditions like COPD, asthma and bronchitis. Explainability aids clinicians in correlating sound patterns with respiratory abnormalities. The integration of these methods not only strengthened the model's transparency but also provided a deeper understanding of how respiratory sound patterns influence classification outcomes. By identifying the most impactful features and analyzing individual predictions, this approach fosters greater confidence in AI-assisted diagnostics. The ability to trace model decisions back to specific acoustic patterns enhances its practical utility, enabling more informed clinical assessments. Moreover, this framework bridges the gap between automated detection and medical expertise, facilitating a more seamless integration of AI tools in respiratory health monitoring.

VII. FUTURE WORK

Building upon the current work, several possibilities for future work can be explored. Mapping the classified respiratory sounds—crackles, wheezes, both or neither—to specific respiratory diseases such as asthma, chronic obstructive pulmonary disease (COPD), or pneumonia. This task will require expanding the current framework and incorporating disease-specific annotations. Additionally, improving the model's accuracy, currently 78%, remains a priority. Advanced deep learning techniques and architectures such as convolutional neural networks (CNNs), which are particularly effective in handling spectrograms of audio data, to better capture complex patterns in respiratory sounds. While this study utilized a stacking model, future study could integrate SHAP (SHapley Additive exPlanations) for more granular feature-level explanations. SHAP being computationally and resource utilization wise intensive, especially for large datasets, making it challenging to implement effectively without GPU support. Combining SHAP with CNN-based models could yield high-performing models bridging the gap between AI-based predictions and clinical decision making. These improvements are more practical and insightful for real-world healthcare applications.

REFERENCES

[1] Chronic obstructive pulmonary disease (COPD) — who.int. [Accessed 30-01-2025].
[2] Norah Saleh Alghamdi, Mohammed Zakariah, and Ha-

nen Karamti. A deep cnn-based acoustic model for the identification of lung diseases utilizing extracted mfcc features from respiratory sounds. Multimedia Tools and Applications, pages 1–33, 2024.

[3] M Atzeni, G Cappon, JK Quint, F Kelly, B Barratt, and M Vettoretti. A machine learning framework for short-term prediction of chronic obstructive pulmonary disease exacerbations using personal air quality monitors and lifestyle data. Scientific Reports, 15(1):2385, 2025.

[4] Gaëtan Chambres, Pierre Hanna, and Myriam Desainte-Catherine. Automatic detection of patient with respiratory diseases using lung sound analysis. In 2018 International Conference on Content-Based Multimedia Indexing (CBMI), pages 1–6, 2018. .

[5] Lobna M Abou El-Magd, Ghada Dahy, Tamer Ahmed Farrag, Ashraf Darwish, and Aboul Ella Hassnien. An interpretable deep learning based approach for chronic obstructive pulmonary disease using explainable artificial intelligence. International Journal of Information Technology, pages 1–16, 2024.

[6] Ambreen Hanif, Xuyun Zhang, and Steven Wood. A survey on explainable artificial intelligence techniques and challenges. In 2021 IEEE 25th international enterprise distributed object computing workshop (EDOCW), pages 81–89. IEEE, 2021.

[7] Reetodeep Hazra and Sudhan Majhi. Detecting respiratory diseases from recorded lung sounds by 2d cnn. In 2020 5th International Conference on Computing, Communication and Security (ICCCS), pages 1–6, 2020. .

[8] Chew-Teng Kor, Yi-Rong Li, Pei-Ru Lin, Sheng-Hao Lin, Bing-Yen Wang, and Ching-Hsiung Lin. Explainable machine learning model for predicting first-time acute exacerbation in patients with chronic obstructive pulmonary disease. Journal of personalized medicine, 12(2):228, 2022.

[9] Apeksha Koul, Rajesh K. Bawa, and Yogesh Kumar. Enhancing the detection of airway disease by applying deep learning and explainable artificial intelligence. Multimedia Tools and Applications, 83(31):76773–76805, Sep 2024. .

[10] Apeksha Koul, Rajesh K Bawa, and Yogesh Kumar. An analysis of deep transfer learning-based approaches for prediction and prognosis of multiple respiratory diseases using pulmonary images. Archives of Computational Methods in Engineering, 31(2):1023–1049, 2024.

[11] Salman Mahmood, Raza Hasan, Saqib Hussain, and Rochak Adhikari. An interpretable and generalizable machine learning model for predicting asthma outcomes: Integrating automl and explainable ai techniques. World, 6(1):15, 2025.

[12] Sara Narteni, Ilaria Baiardini, Fulvio Braido, and Maurizio Mongelli. Explainable artificial intelligence for cough-related quality of life impairment prediction in asthmatic patients. Plos one, 19(3):e0292980, 2024.

[13] Zubaira Naz, Muhammad Usman Ghani Khan, Tanzila Saba, Amjad Rehman, Haitham Nobanee, and Saeed Ali Bahaj. An explainable ai-enabled framework for interpreting pulmonary diseases from chest radiographs. Cancers, 15(1), 2023. ISSN 2072-6694. .

[14] Tahiya Tasneem Oishee, Jareen Anjom, Uzma Mohammed, and Md Ishan Arefin Hossain. Leveraging deep edge intelligence for real-time respiratory disease detection. Clinical eHealth, 2025.

[15] H. Polat and I. Güler. A simple computer-based measurement and analysis system of pulmonary auscultation sounds. Journal of Medical Systems, 28(6):665–672, December 2004. .

[16] Andrea Puig, Miguel Ruiz, Marta Bassols, Lorenzo Fraile, and Ramon Armengol. Technological tools for the early detection of bovine respiratory disease in farms. Animals, 12(19):2623, 2022.

[17] Tabish Saeed, Aneeqa Ijaz, Ismail Sadiq, Haneya Naeem Qureshi, Ali Rizwan, and Ali Imran. An ai-enabled bias-free respiratory disease diagnosis model using cough audio. Bioengineering, 11(1):55, 2024.

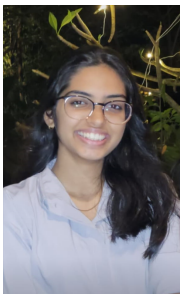
[18] Ahmed I Taloba and RT Matoog. Detecting respiratory diseases using machine learning-based pattern recognition on spirometry data. Alexandria Engineering Journal, 113:44–59, 2025.

[19] Binson VA, M Subramoniam, and Luke Mathew. Non-invasive detection of copd and lung cancer through breath analysis using mos sensor array based e-nose. Expert review of molecular diagnostics, 21(11):1223–1233, 2021.

[20] Xuchun Wang, Yuchao Qiao, Yu Cui, Hao Ren, Ying Zhao, Liqin Linghu, Jiahui Ren, Zhiyang Zhao, Limin Chen, and Lixia Qiu. An explainable artificial intelligence framework for risk prediction of copd in smokers. BMC Public Health, 23(1):2164, 2023.

[21] Yanan Wu, Shuyue Xia, Zhenyu Liang, Rongchang Chen, and Shouliang Qi. Artificial intelligence in copd ct images: identification, staging, and quantitation. Respiratory Research, 25(1):319, 2024.

[22] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. Explainable ai: A brief survey on history, research areas, approaches and challenges. In Natural language processing and Chinese computing: 8th cCF international conference, NLPCC 2019, dunhuang, China, October 9–14, 2019, proceedings, part II 8, pages 563–574. Springer, 2019.



SRUJANA MADIRAJU is currently pursuing her third year B.Tech. degree in Information Technology at Manipal Institute of Technology, Manipal. She is passionate about applying AI to real-world challenges, particularly in healthcare and decision support systems. Her research interests include deep learning, healthcare data analytics, predictive analysis, and artificial intelligence.

MANJULA K SHENOY received her B.Tech. degree in Computer Science and Engineering from Mangalore University in 1994, M.Tech. and Ph.D. in Computer Science and Engineering from Manipal Academy of Higher Education in 2002 and 2014, respectively. She has 30 years of teaching experience and is currently a professor in the Information and Communication Technology Department at Manipal Institute of Technology, Manipal. Her research interests include Semantic Web, Cloud Computing, Data Mining, and Big Data Analytics. She has published over 20 indexed journal papers and 25 conference papers and is currently guiding six Ph.D. scholars.

DHANYA is currently pursuing her undergraduate degree in Occupational Therapy at Manipal College of Health Professions, Manipal Academy of Higher Education, Manipal. Her research interests include respiratory health, rehabilitation sciences, and the application of technology in healthcare to enhance patient outcomes.