SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Multi-Modal Voice Biomarker Analysis For Early Asthma Detection Using Hybrid Deep Learning Models

Kandula Saitharun[1*], T. Srinivasulu[2]

[1*]Department of Electronics and Communication Engineering, Kakatiya University, Warangal, Telangana 506009, India. Email: kandulasaitharun@gmail.com
[2]Department of Electronics and Communication Engineering, Kakatiya University, Warangal, Telangana 506009, India. Email: drstadisetty@gmail.com

Abstract
Early diagnosis of asthma is critical, particularly where access to diagnostic equipment is limited. A non-invasive, intelligent model for early asthma detection is presented, using multimodal voice biomarkers and a hybrid deep learning architecture. Physiological audio cuesphonated vowels, respiratory acoustics, and coughswere recorded from asthma patients and healthy controls, capturing subtle pathological changes in airway function. Preprocessing includes artifact suppression, frequency-domain filtering, time-domain segmentation, and amplitude normalization, followed by extraction of discriminative descriptors: cepstral and spectral representations, perturbation-based voice quality indices, and respiration-specific temporal markers of abnormal airflow. A hybrid convolutional long short-term memory network learns local spectral patterns and temporal dynamics across speech and breathing cycles to distinguish asthmatics from non-asthmatics. Performance is evaluated using accuracy, precision, recall, F1-score, ROC-AUC, and sensitivity. This approach offers a scalable, clinically applicable screening tool with potential in telemedicine, wearable health monitoring, and remote respiratory diagnostics.

Keywords: Deep Learning, Asthma Detection, Multi-Modal Voice Biomarkers, Hybrid CNN–LSTM Model, Respiratory Signal Analysis.

Introduction

Asthma is a persistent inflammatory disease of the respiratory system, which is characterized by millions of individuals around the world and seriously influences the quality of life because of frequent episodes of wheezing, shortness of breath, chest tightness, and cough. Timely diagnosis is important in the successful treatment of diseases, prevention of serious exacerbations, and minimization of respiratory complications in the long term. Nevertheless, traditional methods of diagnosis including spirometry, pulmonary function testing, and physical examination may involve using specialized medical equipment, trained staff, and patient compliance, so they may not be accessible, especially in remote and limited-resource settings (World Health Organization, 2025). This has raised a desire to have more affordable, non-invasive and easy to deploy screening tools to detect asthma at an early age.

The recent progress in speech and respiratory signal processing has provided potential new avenues in identifying diseases through voice-based biomarkers. Respiratory airflow, airway obstruction, and the dynamics of the vocal tract affect human vocalization, and speech, breathing sounds, and cough signals are useful indicators of pulmonary health (Sabry et al., 2024). Subtle acoustic changes in asthmatic people (including abnormal breath patterns, wheezing, abnormal vocal fold vibration and spectral energy distribution) may become early physiological indicators of airway pathology. A combination of these heterogeneous vocal signals offers a more diagnostic representation than a single-modality one analysis.

The suggested paper presents a multi-modal voice biomarker framework of early asthma detection with the help of the latest Deep Learning-based methods. The experiment involves the use of both sustained vowel phonation, respiratory breathing sounds and cough acoustics which are recorded on both the asthmatic and healthy subjects to capture different respiratory features (Petmezas et al., 2022). It uses a preprocessing pipeline to enhance signal quality by removing noise, segmenting and normalizing signals with additional acoustic, phonatory and respiratory features extracted that indicate spectral and temporal abnormalities that indicate asthma.

A hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) architecture is used to model these complex biomarkers well (Abhishek et al., 2024). The convolutional part aims to discover the spatial representations of time-frequency signal patterns automatically and the recurrent part engages the sequential relationship and time variations of respiratory acoustics. This is a hybrid approach to learning that improves classification by integrating the local feature differentiation with the long time temporal modeling.

The proposed study seeks to create a scalable, accurate, and clinically meaningful early screening tool against asthma by combining multi-modal respiratory voice analysis with a hybrid deep learning model (Zhang et al., 2024). This framework is highly applicable in mobile healthcare applications, remote patient monitoring systems, and smart diagnostic support systems to enhance early intervention and increase access to respiratory health testing.

Literature Survey

Saky et al., (2026) suggested a multimodal explainable deep learning and automatic detection of respiratory diseases using audio of the lungs. The architecture is CNNBiLSTM Attention-based spectral-temporal encoder, handcrafted acoustic feature encoder which comprises MFCC, spectral centroid, spectral bandwidth and zero-crossing rate and late-stage fusion to learn other features. With the overall preprocessing techniques of resampling, normalization, noise-filtering and data-augmentation added to it, the model has been able to achieve an accuracy of 91.21 percent, macro F1-score of 0.899 and macro ROC-AUC of 0.98 on the Asthma Detection Dataset Version 2.

The article by Shrivastava et al., (2026) is a review article that carried out the research on speech-based respiratory disorder detection that investigated the possibility of speech analysis as non-invasive diagnostic tool to detect respiratory diseases. The evidence reviewed in the study included peer-reviewed articles published in 2010-2024, which included comparisons of the speech-based method of diagnosis with traditional methods, including lung sound analysis and medical imaging. The review found that there is a gap in research, with most of the current studies mainly covering the lung sounds and imaging and few of them cover speech as an early biomarker of respiratory disorders. It also identified barriers like the non-availability of publicly-available speech data (applied to in only 17% of studies) and the lack of deep-learning use (less than 25% of studies). The results highlight the necessity to have standardized open-source speech databases and more sophisticated machine learning and deep learning models, and it would be possible to combine speech analysis with traditional diagnostic methods to detect and treat respiratory diseases much earlier.

The efficiency of CNNs trained on MFCC features to identify respiratory disorders was proven by Joy and Haider (2021), as they outperformed other traditional machine learning methods in identifying discriminative time-frequency respiratory patterns. Similarly, Kim et al. (2021) used CNN-based models that had pretrained feature extractors to classify abnormal respiratory sounds, with high accuracy in identifying clinically important acoustic events, including crackles and wheezes, and thus demonstrating the potential of deep learning methods in the automatic diagnosis of respiratory diseases.

Vasava and Joshiara (2023) improved the classification of respiratory diseases by converting audio files into the form of 2D spectrograms and using deep CNNs, which proved to be effective in extracting diagnostic acoustic features that were discriminative. Developing this, Hu et al. (2021) improved CNN performance adding depthwise separable convolutions and attention pooling mechanisms, enhancing the robustness of the models and their feature learning capacity. Sfayyih et al., (2021) in their recent systematic reviews have solidified the emerging effectiveness and practicability of deep learning-based acoustic signal analysis in automated lung disease detection, and its possible use in accurate and intelligent respiratory diagnosis.

Demir et al. (2019) investigated the application of pretrained CNNs with conventional classifiers to the analysis of respiratory sounds in spectrograms, and hybrid learning enhances classification results. RDLINet is a lightweight Inception-based CNN architecture proposed by Roy and Satija (2023) that requires fewer computations to detect respiratory diseases. In the same manner, Yadav et al. (2024) trained a CNN-based classifier that had good generalization across pathological and non-pathological respiratory classes, and Wang et al. (2024) introduced a

CNN ensemble with residual attention modules to adequately represent intricate acoustic changes to identify pediatric wheeze. Moreover, copd-specific respiratory sound methods driven by CNN have also demonstrated high levels of discriminative ability to support the argument of deep learning effectiveness to diagnose automated respiratory disease.

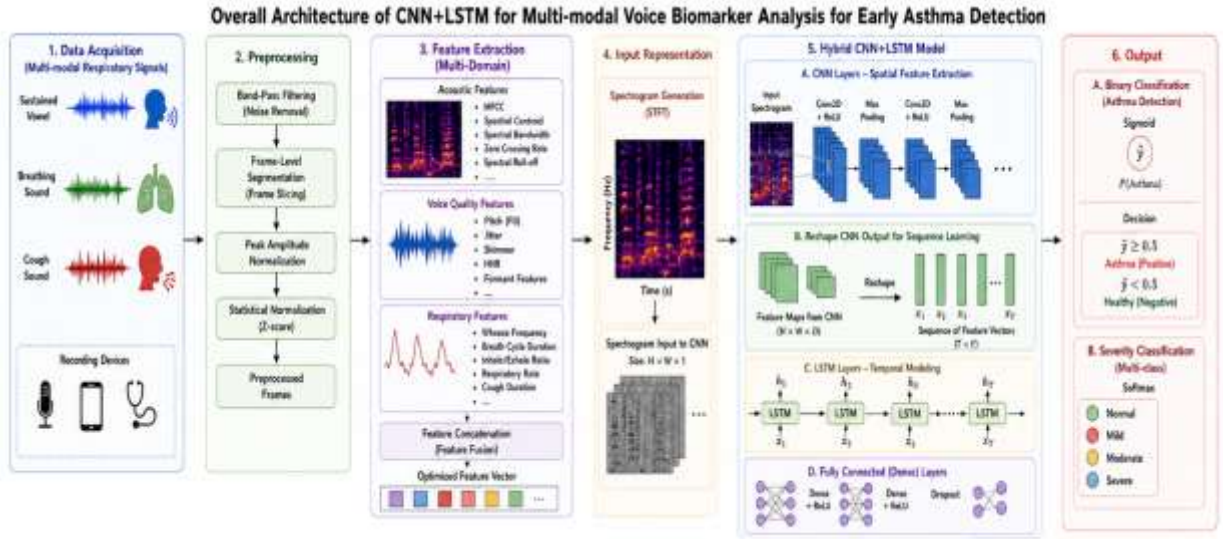
Proposed Work

The proposed work introduces a clever, non-invasive method of the early detection of asthma through the analysis of several voice-related respiratory biomarkers. The essence is that asthma influences airflow patterns, voice tract dynamics and acoustics in respiration and generates changes in speech, breathing sounds, and cough symptoms. Using these modalities in conjunction with state-of-the-art Deep Learning algorithms, the system will assist in delivering precise and accessible early diagnosis.

Pre-Processing

The pre-processing phase is important in the proposed multi-modal voice biomarker analysis to early asthma diagnosis because the quality of the extracted features and predictive potential of the hybrid Deep Learning model mainly rely on the quality of the input respiratory signals (Srivastava et al., 2021). Because the acquired sustained vowel sounds, breathing acoustics, and cough samples might have interfering background, inconsistency in recording, and time lapse, a powerful preprocessing pipeline is required to convert raw audio to standardized and high-quality signals to extract features and classify them. The three key steps in preprocessing in the proposed work are, eliminating noise by band-pass filtering, segmentation of the frame, and normalization.

Figure 1: Overall Architecture of Proposed Work



Noise Removal

The raw respiratory audio signal is very vulnerable to undesired artifacts like environmental noise, hissing of the microphone, electrical noise, reverberation in the room, low-frequency mechanical motion and high-frequency sensor noise.

These noises are capable of concealing delicate biomarkers of asthma problems including wheeze harmonics, cough resonance variation, and the breath-cycle acoustic structures. Hence, a good denoising system is necessary prior to feature extraction.

The use of a Band-Pass Filter (BPF) is based on the fact that the information that is generated by respiration and speaking falls within a certain frequency range, whilst irrelevant frequencies outside of the range can be safely eliminated.

Assume that the raw audio signal that has been recorded is: x(t) versus time index t, then the filtered signal is:

$$y(t) = x(t) * h(t) \quad (1)$$

Where,

y(t) filtered signal, h(t) impulse response of band-pass filter and $*$ convolution operator.

The ideal band-pass filter transfer function is.

$$H(f) = \begin{cases} 1, & f_L \le f \le f_H \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

Where,

f_L cutoff frequency and f_H cutoff frequency.

Adaptive band-pass filtering in the proposed respiratory signal processing framework is done by choosing the frequency ranges depending on respiratory signal type. Some speech and sustained vowel sounds are applied in the range between 80 and 400 Hz, breathing sounds in the range between 100 and 2500 Hz, wheeze components in the range between 100 and 1600 Hz, and cough acoustics in the range between 50 and 5000 Hz. Such a signal-specific cutoff choice contributes to the protection of significant respiratory biomarkers, noise reduction, and higher accuracy of feature extraction towards the detection of asthma.

Segmentation

Segmentation is a basic preprocessing phase that breaks down continuous respiratory audio recordings into small parts that can be analyzed, known as frames. As sustained vowel sounds, breathing acoustics and cough signals are non-stationary bioacoustic signals, the statistical and spectral characteristics of these signals change continuously with time (Saky and Tshering, 2025). Symptoms like wheezing, coughing in brief bursts, airflow obstruction patterns, and irregular breathing cycles might only be short-lived. Thus, localized pathological variations may be buried in the whole recording as a single signal. Frame-level slicing is a solution to this problem, as it divides the signal into short-period segments, allowing detailed analysis of their time localization and enhances the ability to extract features and further Deep Learning model representation.

Let filtered signal be y(n) now L and R are length and shift.

The frames can be determined by consideration of

$$K = \frac{N-L}{R} + 1 \quad (3)$$

For each frame $y_k[m] = y[m + kR], 0 \le m \le L$

Where,

kis a frame number and *m*indicated the sample index in the frame.

Frame Duration Selection

The frame duration is chosen depending on the sampling frequency f_s.

$L = T_f X f_s$ (4)

Where,

T_f is the duration of the frame in seconds and f_s is the sampling frequency.

For example, if $f_s = 16000Hz$ and frame duration is 25 ms:

$L = 0.025X16000 = 400\ samples$

$R = \frac{L}{2} = 200$ (5)

This overlap offers a continuous change of time and removal of loss of transient respiratory events between frame boundaries.

Each frame is multiplied by a window function to minimize spectral leakage and discontinuities at frame edges to reduce discontinuities at frame edges.

Normalization

Normalization is a necessary preprocessing step in the suggested multi-modal voice biomarker analysis framework to detect early onset asthma, as it makes sure that sustained vowel sounds, breathing acoustics, and cough recordings can be matched to a common numerical scale prior to feature extraction and a deep learning-based classification (Kapetanidis et al., 2024).

In the case of the proposed asthma detection system, a hybrid normalization method is suggested that involves peak normalization to equalize the amplitude of recording and Z-score normalization to statistically normalize it. This mixed approach provides consistency of amplitude, statistical consistency and reliability of features. Table 1 summarize the steps involved in the pre-processing.

Table 1: Steps involved in Pre-Processing

Step No.	Preprocessing Step	Method	Description	Purpose / Outcome

1	Noise Removal	Band-Pass Filtering	Use of a frequency-selective filter to pass useful respiratory and vocal frequency content and reject low-frequency drift, environmental noise, microphone artifacts and high-frequency interference of sustained vowel, breathing and cough records.	Improves signalto-noise-ratio (SNR), preserves respiratory acoustic features (wheeze harmonics, cough resonance), and better downstream feature extraction consistency.
2	Frame Slicing	Frame Level Segmentation	Splits the continuous filtered respiratory audio signal into brief overlapping frames of constant length (usually 20-40 ms) to allow local short-time analysis of non-stationary acoustic events. The purpose of windowing functions like Hamming window is to reduce the edge discontinuities.	Records breathe-by-breathe changes, separates inhalation/exhalation, allows frame-by-frame capture of acoustic biomarkers, and offers sequential information in a structured format to train CNNLSTM temporal learning.
3	Normalization	Peak Amplitude	Scales every segmented frame according to the largest absolute amplitude in the frame, such that signal values are converted to a standard amplitude range (usually -1 to 1).	Eliminates amplitude differences due to recording loudness, microphone placement, and device sensitivity, and makes the amplitudes of all respiratory recordings consistent.
4		Z-Score Standardization	Normalizes every normalized frame by its mean, and by its standard deviation, making the mean value zero and the unit variation one.	Minimizes inter-subject variance, enhances numerical accuracy, avoids bias during feature extraction and the hybrid Deep Learning model converges more quickly and reliably.

Feature Extraction

The proposed work is predicated on the fact that efficient feature extraction strategy is necessary to transform preprocessed respiratory audio signals into useful numerical features that can be intelligently classified. As asthma affects the airflow of the respiratory system, the phonatory mechanism, airway resonance, and breathing rhythm in parallel, it is impossible to describe variations in the disease in just one category of features (Xu et al., 2026). The proposed work attempts to overcome this limitation by using a Multi-Domain Feature Extraction Method, where discriminative biomarkers are systematically obtained using three complementary domains, i.e., acoustic domain, voice-quality domain, and respiratory domain. This combined representation helps the hybrid Deep Learning model to learn more delicate pathological patterns related to the early-stage asthma. For easy under the feature extraction is summarized in Table 2.

Acoustic Domain Feature Extraction

The acoustic domain models frequency-dependent, spectral, and temporal changes in energy in respiratory signals. They are effective especially in detecting airflow turbulence, cough resonance abnormalities, and subtle spectral changes as a result of airway constriction.

Each segmented frame is subjected to Short-Time Fourier Transform (STFT) to study short-time acoustic behavior:

$$X(m,w) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-jwn} \quad (6)$$

Where $x[n]$ is the input at a respiratory frame, $w[n]$ analysis window, m frame position, w angular and $X(m,w)$ spectral representation. This produces a timefrequency map, which shows the local respiratory acoustic behavior.

Mel Frequency Cepstral Coefficients (MFCC)

MFCCs are a compact representation of the spectral envelope in terms of human auditory perception. Frequency scale is converted to Mel scale:

$$Mel(f) = 2595\, log_{10}\, 1 + \frac{f}{700} \quad (7)$$

Cepstral coefficients are derived with the help of Discrete Cosine Transform after Mel filterbank processing and logarithmic compression:

$$C_k = \sum_{m=1}^{M} log(S_m)\, cos\, \frac{\pi k}{M}(m - 0.5) \quad (8)$$

Where,

S_m Mel filterbank energy and M number of Mel filters, C_k kth MFCC coefficient.

MFCCs represent successfully respiratory spectral variations that can be ascribed to wheezing and constricted airways.

Spectral Centroid

The centroid is the point of gravity of spectral energy:

$$SC = \frac{f|X(f)|}{|X(f)|} \quad (9)$$

An increase in centroid values reflects more energy in the high frequency, which is likely to be a wheeze component.

Zero Crossing Rate (ZCR)

Zero Crossing Rate is used to measure the rapid polarity changes:

$$ZCR = \frac{1}{2N}\sum_{n=1}^{N} |sign(x[n]) - sign(x[n-1])| \quad (10)$$

ZCR helps to describe noisy respiratory bursts and cough sharpness. So that acoustic qualities are:

$$F_A = [MFCC, SC, ZCR] \quad (11)$$

Voice-Quality Domain Feature Extraction

Asthma changes the respiratory support in the process of phonation, which influences the consistency of vibration of vocal folds and the stability of the voice. Such physical changes are quantified during the voice-quality assessment.

Fundamental Frequency (Pitch)- Pitch is determined based on the period of glottal vibration: $F_0 = \frac{1}{T_0}$, T_0 is the vocal period.

Jitter- Jitter is a measure of cycle-to-cycle frequency perturbation:

$$Jitter = \frac{1}{N-1}\sum_{i=1}^{N-1} |\frac{T_i - T_{(i+1)}}{T'}| X100 \quad (12)$$

Where,

T_i pitch period and T' is average pitch period. Increased jitter implies wobbly vocal vibration.

Shimmer- Shimmer is a measure of amplitude perturbation:

$$Shimmer = \frac{1}{N-1}\sum_{i=1}^{N-1} |\frac{A_i - A_{i+1}}{A'}| X100 \quad (13)$$

A cycle amplitude A_i and A' mean amplitude. High shimmer is an indicator of uneven airflow pressure.

The final quality vector is $F_v = [Pitch, Jitter, Shimmer, \ldots\ldots]$

Respiratory Domain Feature Extraction

Biomarkers in the respiratory domain are a direct description of the breathing mechanics and patterns of airway obstruction.

Wheeze Energy Detection

Wheeze is located in high-frequency narrow bands. The Wheeze energy ratio is calculated by using:

$$WER = \frac{\sum_{f=f_1}^{f_2} |X(f)|^2}{\sum_{f=0}^{f\_max} |X(f)|^2} \quad (14)$$

Where,
f1and f2 wheeze frequency band.

Breath Cycle Duration
Complete breathing cycle:

$$T_b = T_{inhale} + T_{Exhale} \quad (15)$$

Prolonged expiration is common in asthmatic subjects:

$$\frac{T_{Exhale}}{T_{inhale}} > 1 \quad (16)$$

This ratio becomes a strong biomarker.

Respiratory Rate
Breathing rate is the frequency of breathing defined as:

$$RR = \frac{60}{T_b} \quad (17)$$

Abnormal respiratory rate can be a sign of airway dysfunction.
Respiratory vector:

$$F_R = [WER, T_b, RR, \dots.] \quad (18)$$

Multi-Domain Feature Fusion
The characteristics of all domains are concatenated:

$$F_{fusion} = F_A \oplus F_v \oplus F_R \quad (19)$$

and $\oplus$ is concatenation of vectors, Where Acoustic = p features, Voice = q features, Respiratory = r features. Then, $Dim(F_{fusion}) = p + q + r$.

Feature Optimization
To remove redundancy:

$$Score_i = \frac{Var(F_i)}{Var(F_j)} \quad (20)$$

Very informative features are stored and dimensionality is reduced without the loss of the discriminative power.
Final optimized vector:

$$F^* = [f_1, f_2, \dots. f_k] \quad (21)$$

This streamlined multi-domain biomarker vector is then sent to the CNN-LSTM classifier.

Table 2: Steps in Feature Extraction

Step No.	Feature Extraction Step	Method	Description	Purpose / Outcome
1	Short-Time Spectral Analysis	Short-Time Fourier Transform (STFT)	Coded respiratory audio to timefrequency representation, and spectral analysis of dynamic respiratory events locally.	Allows detection of temporary breathing phenomena like wheezing, cough bursts and acoustic changes in the breath cycle.
2	Acoustic Domain Extraction	MFCC (Mel Frequency Cepstral Coefficients)	Gets perceptually significant cepstral features of respiratory spectral patterns.	Records small variations in frequency envelopes that are related to airway obstruction, cough resonance, and distorted respiratory acoustics.
		Spectral Centroid	A spectral value which shows the center of gravity of frequency distribution of a	Measures with the largest amount of spectral energy in the respiratory audio.

			sound signal.	
		Zero Crossing Rate (ZCR)	An attribute of time that quantifies the rate of change of the signal waveform of sign (or crosses the zero amplitude axis).	Measures the signal noisiness and fluctuations of waveforms in respiratory recordings.
3	Voice-Quality Domain Extraction	Fundamental Frequency (Pitch)	Determine the frequency of vibration of the vocal folds in prolonged phonation.	Determines the instability of pitch and phonatory changes due to the lessening of respiratory support in asthmatic subjects.
		Jitter	A voice perturbation measure that is a measure of cycle to cycle differences in vocal frequency.	Assesses the periodicity of vocal fold vibration.
		Shimmer	A parameter of voice quality that quantifies cycle to cycle change in vocal intensity.	Measures stability of amplitude with sustained phonation.
4	Respiratory Domain Extraction	Wheeze Energy Ratio (WER)	A respiratory acoustic index that approximates the ratio of signal energy the frequency bands of wheeziness.	Measures the intensity and prevalence of wheezing in the recordings of breaths.
		Breath Cycle Duration	Fractions the overall respiratory cycle.	Helps in determine long expiration, abnormal breathing rhythm, and disrupted patterns of respiratory timing of asthma.
		Respiratory Rate	Calculates the rate of breathing in minutes.	Identifies abnormal breathing and respiratory pacing abnormalities associated with airway dysfunction.
5	Feature Fusion	Multi-domain feature concatenation	Combines acoustic, voice-quality and respiratory biomarkers into a single feature vector.	Generates a multi-dimensional respiratory representation that is more comprehensive and enhances disease characterization and classification performance.
6	Feature Optimization / Selection	Variance-based feature scoring	Determines the feature importance and eliminates redundant features	Enhances the discriminative ability, computation efficiency, and minimizes the complexity of the models prior to the classification.
7	Final Optimized Feature Vector	Selected multi-domain biomarker vector	Processes the last input to CNN-LSTM classification.	Offers a small, powerful, and very discriminant biomarker panel to identify early asthma.

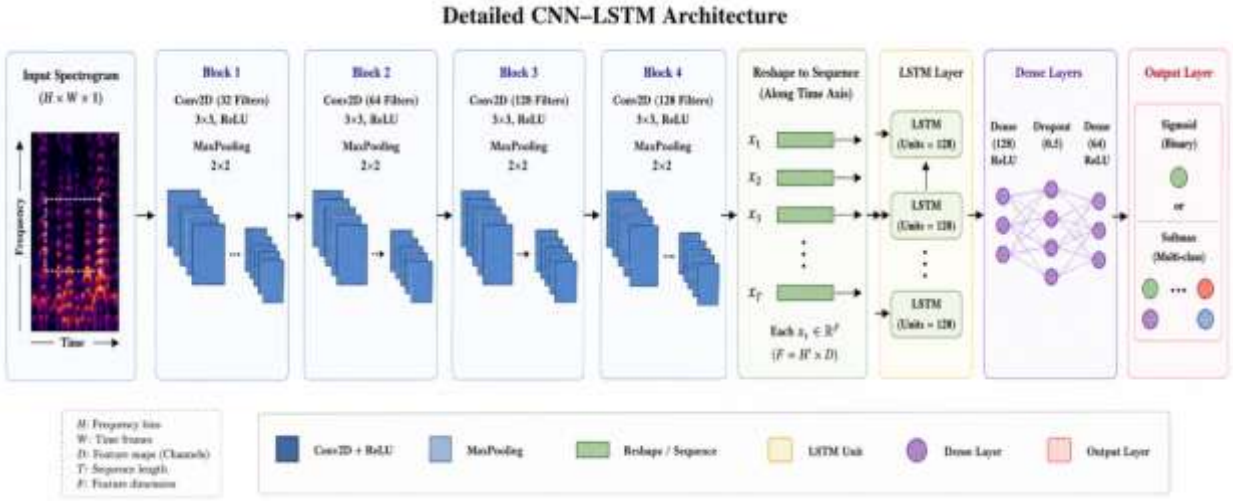
3.3. Classification

In the suggested work the classification step is carried out on a hybrid of Convolutional Neural Network-Long Short-Term Memory (CNNLSTM) model. This choice of a hybrid architecture is driven by the fact that respiratory bioacoustic signals are very complex, with spatial-frequency patterns and time-dependent breathing dynamics being potentially associated with important diagnostic information. The discriminative biomarkers found in sustained vowel sounds, breathing acoustics and cough signals include acoustic irregularities of wheeze, spectral anomalies, vocal perturbations, and breathing cycle variations that change over time. These multi-dimensional dependencies may not be well-represented by a traditional machine learning-based classifier. Thus, the suggested work is a combination of CNN to learn the deep spatial features and LSTM to learn the sequence of time models, and a fully connected classification layer is used to end with the final disease prediction.

The optimized multi-domain set of features based on acoustic, voice-quality, and respiratory domains is transformed to a time-frequency spectrogram representation, which is input to the hybrid deep learning model.

The respiratory audio signals (sustained vowel, breathing sound, cough sound) undergo multi-domain feature extraction and preprocessing followed by transforming them into time-frequency spectrogram representations. A spectrogram graphically represents the distribution of signal energy with time and frequency, and retains detailed acoustic information about respiration.

Figure 2: Architecture of CNN for Proposed Work



CNN Layer Processing

The CNN is the initial step of the hybrid model and conducts hierarchical spatial feature extraction of spectrograms. Its primary goal is to detect automatically the wheeze bands, cough harmonic patterns, spectral shapes, breathing energy patterns, turbulence patterns, and abnormal resonance patterns.

Convolution Operation

The CNN uses a small learnable matrix which is called a kernel or filter represented as:

$$K \in R^{m\,X\,n} \qquad (21)$$

where $m\,X\,n$ is the size of the receptive field, usually 3 x 3or 5 x 5. All kernels have trainable weights, which modify during model training to identify certain respiratory patterns.

The convolution operation is mathematically expressed as:

$$F(i,j) = \sum_m \sum_n I(i-m, j-n)K(m,n) \qquad (22)$$

Where,

$I(i,j)$ is spectrogram pixel, $K(m,n)$ filter weights and $F(i,j)$ is output feature map.

Rather than using one kernel, multiple convolution filters can be used at the same time like $\{K_1, K_2, \ldots\ldots, K_N\}$.

In the suggested CNN architecture, the convolutional filters are trained to extract a certain breathing acoustic pattern in the input spectrogram. Wheezing, cough spectral edges, breathing harmonic textures, and airflow turbulence signatures are some of the different biomarkers which are automatically identified by different filters. This feature learning is specialized and allows the model to learn the various respiratory abnormalities to enhance the accuracy of asthma detection.

This generates a series of feature maps, $F = \{F_1, F_2, \ldots\ldots F_N\}$which accumulate to create a profound respiratory representation.

Non-Linear Activation

The result of the convolution layer is fed into a non-linear activation function and then sent to another processing unit. The main idea of non-linear activation is to add mathematical non-linearity to the network to allow the Convolutional Neural Network (CNN) to achieve the complex respiratory acoustic relationships that could not be captured with simple linear transformations. The respiratory bioacoustic cues which are sustained vowel phonation, breathing sounds and cough acoustics are highly nonlinear with changes in airway obstruction, wheeze intensity, airflow turbulence and respiratory rhythm abnormalities. As such, to learn subtle asthma-related biomarkers using spectrogram representations, it is necessary to include non-linearity.

When there were only convolution operations, the whole CNN would act as a linear system regardless of how many layers were stacked. The linear transformation can be expressed mathematically as:

$$y = Wx + b \quad (23)$$

Where,

x is the input feature, W is the weight matrix, b is the bias term and y is the response.

The Rectified Linear Unit (ReLU) will be utilized in the proposed work because it is computationally efficient and it performs better in learning deep features.

In the proposed work, the Rectified Linear Unit (ReLU) is employed due to its computational efficiency and superior performance in deep feature learning.

The activation function of the ReLU is given as:

$$ReLU(x) = max(0, x) \quad (24)$$

The activation function transforms negative responses to zero, and maintains positive activations, enabling the network to highlight significant respiratory biomarkers, and deadweight or irrelevant signals. This step improves the representation of significant features, sparsely activates and reduces the complexity of the computations, which improves the efficiency and learning capacity of the proposed CNN model in detecting asthma.

When using the convolution output $F(i, j)$ the activated feature map is.

$$A(i, j) = ReLU(F(i, j)) \quad (25)$$

$F(i, j)$ is convolution output and $A(i, j)$ activated output.

Pooling Layer

After the convolution and non-linear activation, the second most important step in the proposed CNN architecture is the Pooling Layer. Pooling is a dimensionality reduction operation that reduces feature map compressing, but does not eliminate the most informative respiratory characteristics learned during convolution. The main goal of pooling is to minimize the computational complexity, enhance feature resiliency, and store dominant respiratory signatures related to asthma to further process them. The feature map is represented as after activation as:

$$A \in R^{(H\ X\ WXD)} \quad (26)$$

where:

H = height of feature map, W = width of feature map and D = features/channels.

Max Pooling is used in the proposed work as it maintains the best respiratory activation in each local region, which has the great advantage of identifying localized pathological respiratory biomarkers. The mathematical meaning of max pooling is max pooling :

$$P(i, j) = \max_{(u,v)\in R} A(u, v) \quad (27)$$

Where,

R denotes local pooling region, $A(u, v)$ denotes values of activation within region R, $P(i, j)$ denotes pooled output. This operation only selects the maximum activation in the pooling window.

Reshaping CNN Output for Sequence Learning

In respiratory spectrograms, time progression is plotted on the horizontal axis, and frequency components on the vertical axis. As LSTM models time, the CNN feature tensor is reorganized along the time axis in such a way that each time slice becomes a single sequential input vector. CNN output tensor $Z_{CNN}(H, W, D)$ is converted to $X_{Seq}(W, HXD)$. This is to say that the width dimension W is the consecutive time step and the height and channel dimensions are reduced to a single feature vector.

Therefore, the reconstruction of CNN output to sequence learning is the intermediary process between the spatial feature learning and the temporal sequence learning of respiratory biomarkers, which allows the hybrid CNN-LSTM architecture to learn and predict both the fixed and dynamic respiratory biomarkers to identify early asthma correctly.

LSTM Layer Processing for Temporal Learning

Following CNN feature extraction and reshaping, the LSTM input will be $X_{seq} = [x_1, x_2, \ldots . x_T]$ with each x_t having deep respiratory content at a specific time instant, including learned wheeze pattern patterns, cough acoustic pattern, breathing harmonic texture, and airflow spectral pattern. The purpose of LSTM is to investigate the manner in which such respiratory characteristics change over time.

The LSTM layer in the proposed model has two internal states: the hidden state h_t, the state that holds the short term sequential respiratory information, and the cell state C_t, the state that holds long term temporal memory. The combination of these states enables the network to selectively memorize significant breathing patterns and forget irrelevant information to learn effectively the temporal breathing dynamics in the accurate asthma detection.

Forget Gate

The forget gate is the initial step in LSTM and it dictates what information is to be retained or forgotten in the previous memory state. Not all short time variations in respiratory sequence analysis are clinically significant; there may be environmental noise or recording artifact, or the normal variation in breathing. Such irrelevant information is filtered out by the forget gate.

The forget gate is defined as: $f_t = \sigma\big(W_f[h_{t-1}, x_t] + b_f\big)$, the sigmoid generates values $0 \leq f_t \leq 1$. Based on this values the gate chooses value close to 0 to forget the information and value close to 1 to retain the information. This enables the model to eliminate irrelevant respiratory variations whilst retaining significant asthma-associated time biomarkers.

Input Gate – Learning New Respiratory Information

The input gate decides the degree to which it should store in the memory new respiratory information at the current time step. This allows LSTM to revise its knowledge when new pathological patterns appear, e.g. the development of wheezing, bursting cough, or change in the rhythm of breathing.

Cell State Update – Long-Term Respiratory Memory

The long term memory of the LSTM is the cell state. It retains clinically significant respiratory patterns over a period of time. The update equation is:

$$C_t = f_t \odot C_{t-1} + i_t \odot C_t' \quad (28)$$

Where,

C_{t-1} is past cell memory, $f_t \odot C_{t-1}$ is stored respiratory memory, $i_t \odot C_t'$ newly added respiratory information and $\odot$ element-wise multiplication.

This memory build-up is very much significant as the symptoms of asthma are not that much manifested in individual respiratory incidents but on a long term pattern. In this way cell state can store long exhalation patterns, recurring wheezes patterns, cumulative cough bursts, sustained breathing abnormalities.

Output Gate – Producing Temporal Respiratory Representation

The output gate is used to decide what data in the updated memory is to be passed on as the hidden representation in the next time step.

Output gate activation:

$$O_t = \sigma(W_o[h_{t-1}, x_t] + b_0 \quad (29)$$

Hidden state generation:

$$h_t = O_t tanh(C_t)$$

Where,

O_t regulates the strength of output and h_t is the time respiratory representation at time t.

The repressed state is full of serial respiratory knowledge, acquired up to that point.

Fully Connected or Dense layer

Once the CNN has learned the discriminative spatial respiratory biomarkers and the LSTM network has learned the dynamics, the resulting high-level feature representation will have to be transformed into a useful clinical prediction. It is the dense layer that does this transformation and sums up all the obtained respiratory characteristics and recreates them into a classification space to determine whether the subject is an asthmatic or a healthy individual. The discriminative engine is then the fully connected layer which transforms deep learned representations to diagnostic outputs.

Input to the Dense Layer

After modeling time sequences using the LSTM, the result is either represented as the last hidden state or as a summary in the form of a sequence:

$$H = [h_1, h_2, \ldots . h_T] \quad (30)$$

Where,
h_t hidden representation at time step t and T total sequence length.
In most applications, the last hidden state h_T or a pooled encoding of all hidden states is chosen as the compact encoding of the entire respiratory sequence. This vector holds the coded information about persistence of wheezing, cough sequence features, irregularity of breathing rhythm, airflow obstruction signature and pattern of phonatory instability. It is this learned feature vector that is input to the dense layer.

Linear Transformation
The dense layer uses a weighted linear transformation of the input feature vector. All the neurons of the dense layer are connected to all the neurons of the last layer, hence this makes it fully connected. Mathematically, the operation is given as follows:

$$z = Wx + b \quad (31)$$

Where,
x = input feature vector from LSTM,
W = learnable weight matrix,
b = bias vector,
z = transformed output vector.
This metamorphosis transfers the acquired respiratory code to a novel feature space whereby the patterns of classes become increasingly distinguishable.
Individual neurons in the dense layer learn a weighted sum of respiratory biomarkers with more weight given to clinically significant features like persistent wheeze patterns, long expiration patterns, repeated cough bursts, or distorted breathing rhythm patterns.

Nonlinear Decision Learning
In order to enhance the representational power, the transformed output is usually subjected to an activation function:

$$a = \emptyset(z) \quad (31)$$

where:
a = activated dense representation,
$\emptyset(z)$ = activation function such as ReLU.
Nonlinear activation used enables the dense layer to represent nonlinear interactions between respiratory biomarkers and not just linear separability.

Hidden Dense Layer
Hidden dense layers can be used in deeper architectures:
$a^{(1)} = \emptyset(W^{(1)}x + b^{(1)}$ and $a^{(2)} = \emptyset(W^{(2)}x + b^{(2)}$.
The intermediary thick layers increasingly hone the abstraction of respiratory features and enhance the discrimination of classes.

Output Classification Layer
The last thick layer generates the classification output. A single output neuron with Sigmoid activation is normally employed in binary asthma detection:

$$y' = \frac{1}{1+e^{-z}} \quad (32)$$

Where, y' is the predicted probability of asthma. The format of decision rule can be Suppose, $y' \geq 0.5$ is forecasted as Asthma Positive and $y' < 0.5$ is forecasted as Healthy.

Experimental Results
Experimental testing of the proposed CNNLSTM framework illustrates a promising result in detecting asthma at its early stages with the help of multi-modal respiratory voice biomarkers. The hybrid architecture is able to learn spatial acoustic signature and temporal respiratory dynamics and thus improves on classification ability. The hyper parameter of the proposed CNN-LSTM is indicated in the Table 3.

Table 3: Hyperparameter Settings

Hyperparameter	Value
Input Size	128 X 128 X1

CNN Block 1 Filters	32
CNN Block 2 Filters	64
CNN Block 3 Filters	128
CNN Block 4 Filters	128
Kernel Size	3X3
Stride	1
Activation Function	ReLU
Pooling Type	MaxPooling
Pooling Size	2X2
LSTM Layers	2 Stacked Layers
LSTM Units	128 per layer
Dropout Rate	0.5
Dense Layer 1	128 neurons
Dense Layer 2	64 neurons
Output Layer	1 neuron (Sigmoid) / Multi-class Softmax
Optimizer	Adam
Learning Rate	0.001
Batch Size	32
Epochs	50-100

Dataset

Asthma Detection Dataset: Version 2 Overview Asthma Detection Dataset: Version 2 is a specialized set of audio samples to support machine learning (ML) and deep learning (DL)-based research in diagnosing asthma and other lung diseases based on kaggle-collected samples. This is a self-generated, well-managed dataset that contains sound files broken down into 1.5-5 seconds to be a consistent unit and easy to work with in analysis. The dataset is the perfect one to build and test ML models that can detect asthma by analyzing lung sounds. The dataset is a sum total of 1,211 audio samples that are categorized into five different classes based on different lung conditions:
Asthma: 288 samples
Bronchial: 104 samples
COPD: 401 samples
Healthy: 133 samples
Pneumonia: 255 samples

Evaluation Metrics

In the suggested CNN-LM-based multi-modal voice biomarker analysis system to identify early asthma, the model performance is assessed with the help of the typical classification measures that assess the efficiency of distinguishing between asthmatic and healthy individuals. The confusion matrix is based on these measures and is determined by:
True Positive (TP): Asthmatic subjects that are asthma positive.
True Negative (TN): Healthy individuals who were correctly judged to be healthy.
False Positive (FP): Healthy individuals that were incorrectly diagnosed to be asthma positive.
False Negative (FN): Asthmatic patients who were misdiagnosed as healthy.

Accuracy

The measure of accuracy is the proportion of all the cases that are correctly identified of all the samples that are tested. It assesses the degree to which the model differentiates asthmatic and healthy subjects, as a group.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (33)$$

Sensitivity
Sensitivity is used to determine the rate of correctly detected actual asthma cases by the model. It assesses the model in terms of positive asthma conditions detection.

$$Sensitivity = \frac{TP}{TP+FN} \quad (34)$$

Recall
Recall is the ratio of the amount of relevant positive cases that are recalled by the classifier. Sensitivity equals mathematically recall in binary medical diagnosis.

$$Recall = \frac{TP}{TP+FN} \quad (35)$$

F1-Score
The harmonic mean of precision and recall is F1-score. It gives a balanced measure where the distribution of classes is not balanced.

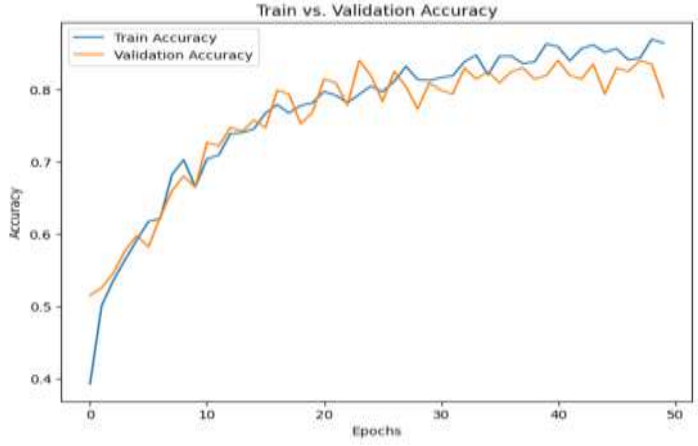
$$F1 = 2X\frac{Precision\ X\ Recall}{Precision+Recall} \quad (36)$$

ROC–AUC
ROC-AUC is a measurement that could be adopted to quantify the ability of the classifier to distinguish between negative and positive classes at different classification thresholds. It is a sensitivity/false positive rate plot.

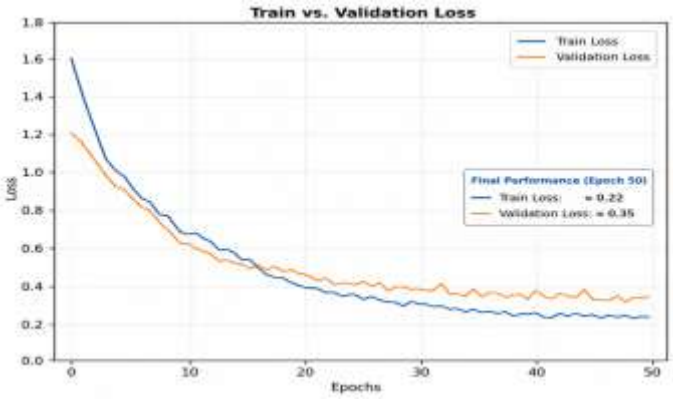
Quantitative Performance Analysis

Accuracy Analysis
The accuracy graph of the training and validation in Figure 3 shows that the proposed CNNLSTM model is well learned and converges smoothly with the training epochs. Both accuracy curves demonstrate the steady improvement with the training accuracy being about 86-87 percent, and validation accuracy converging to 82-84 percent. The slight distance between the two curves indicates that there is good generalization and that there is no overfitting. All in all, the model has an estimated classification of 84-86, which is good at detecting early asthma using multi-modal voice biomarkers.

Figure 3: Accuracy analysis in Training vs Validation



Loss Analysis
As indicated by the training and validation loss curve in Figure 4, the proposed CNN-LSTM model converges well with steady learning during training. The training loss drops to 0.22 and the validation loss drops to 0.35 signifying that it is optimizing successfully and prediction ability has improved. The narrow separation between the two curves indicates that there is good generalization with less overfitting. The model in general shows strong learning in terms of proper early asthma detection.
Figure 4: Loss Analysis in Training Vs Validation



Confusion Matrix

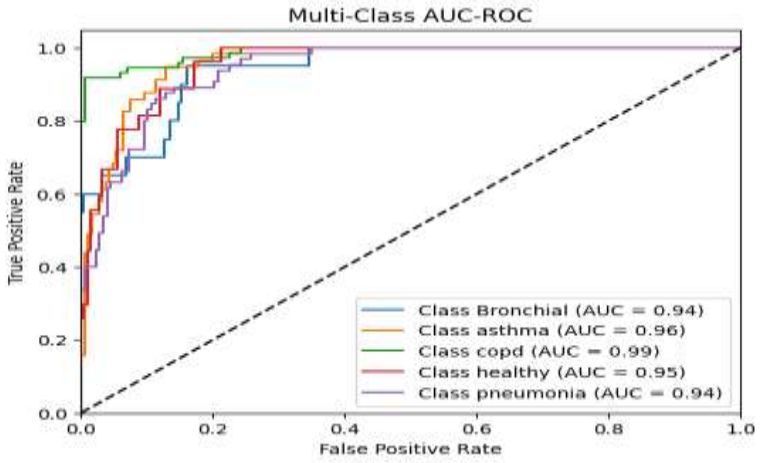
The confusion matrix in Figure 5 shows that the proposed CNN-LSTM model has high multi-class classification accuracy, and most predictions are located along the diagonal, which implies accurate class recognition. The model works remarkably well with Class 2 (71 correct predictions), Class 1 (50 correct) and Class 4 (42 correct) with a high level of discriminative power of these groups. Class 0 and Class 3 exhibit moderate misclassifications implying the existence of some overlapping respiratory patterns. In general, the confusion matrix results show that the proposed framework is effective and reliable when it comes to classifying early asthma detection, whereas the results reveal that additional work can be done to decrease the frequency of the inter-class confusion.

Figure 5: Performance analysis of Confusion Matrix

ROC-AUC Curve

As shown in the multi-class AUC-ROC graph in Figure 6, the proposed CNN-LSTM model demonstrates high classification performance in all respiratory classes, with ROC curves near the upper-left corner, which represents high sensitivity and low false positive values. The model indicates the optimal discriminatory performance on COPD with an AUC of 0.99, followed by Asthma (0.96) and Healthy (0.95) and Bronchial (0.94) and Pneumonia (0.94) also indicate strong classification performance. Because the complete set of class-wise AUCs is greater than 0.90, the model is highly robust and class separable with good predictive accuracy in the differentiation of various respiratory diseases, which confirms the suitability of the proposed framework to accurately classify respiratory diseases.

Figure 6: AUC-ROC curve



Conclusion

The research introduced a new multi-modal voice biomarker study framework to detect asthma at an early stage with the hybrid CNN-LSTM deep learning model. The suggested system is useful in combining long term vowels sounds, breathing acoustics, and cough indicators to record various respiratory biomarkers related to asthma. A preprocessing pipeline, which included band-pass filtering, frame-level segmentation and normalization, was used to improve the quality of the signal and guarantee accurate feature extraction. Moreover, the implementation of a multi-domain feature extraction algorithm, which integrated the use of acoustic, voice-quality, and respiratory biomarkers, made it possible to provide a full description of respiratory abnormalities. The CNN element was good enough to acquire discriminative spatial spectrogram characteristics whereas the LSTM element was good enough to acquire temporal respiratory characteristics such that a solid framework was formed that could capture both static and dynamic disease patterns.

The experimental analysis showed that the proposed model had a high classification performance with a training accuracy of about 86-87, validation accuracy of about 82-84 and a steady reduction in the training and validation loss values, which proved stable convergence and good generalization with a low overfitting. The confusion matrix also verified the robust multi-class classification accuracy of about 77.8% with multi-class AUC-ROC values of over 0.90 at all respiratory classes indicating good discriminative and strong predictive power. These results confirm that multi-modal respiratory voice biomarkers with hybrid deep learning is an effective, non-invasive, and clinically meaningful method of detecting the presence of asthma early allowing it to be used in smart healthcare systems, telemedicine, and remote respiratory monitoring systems with a lot of potential.

Acknowledgement
None

Data availability statements
Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Funding
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of Interest
The authors declare that they have no potential conflict of interest.

References

1. Abhishek, S., Ananthapadmanabhan, A., Anjali, T., Reyma, S., Perathur, A., & Bentov, R. B. (2024). Multimodal integration of an enhanced novel pulmonary auscultation real-time diagnostic system. IEEE MultiMedia, 31(3), 18–43.
2. Demir, F., Sengur, A., & Bajaj, V. (2019). Convolutional neural networks based efficient approach for classification of lung diseases. Health Information Science and Systems, 8(1), 4.
3. Hu, G., Wang, K., & Liu, L. (2021). Underwater acoustic target recognition based on depthwise separable convolution neural networks. Sensors, 21(4), 1429.
4. Joy, R. P., & Haider, N. S. (2021). Classification of lung disorders based on lung sounds using deep learning. Journal of Physics: Conference Series, 1937(1), 012036.
5. Kapetanidis, P., Kalioras, F., Tsakonas, C., Tzamalis, P., Kontogiannis, G., Karamanidou, T., Stavropoulos, T. G., & Nikoletseas, S. (2024). Respiratory diseases diagnosis using audio analysis and artificial intelligence: A systematic review. Sensors, 24(4), 1173.
6. Kim, Y., Hyon, Y., Jung, S. S., Lee, S., Yoo, G., Chung, C., & Ha, T. (2021). Respiratory sound classification for crackles, wheezes, and rhonchi in the clinical field using deep learning. Scientific Reports, 11(1), 17186.
7. Petmezas, G., Cheimariotis, G.-A., Stefanopoulos, L., Rocha, B., Paiva, R. P., Katsaggelos, A. K., & Maglaveras, N. (2022). Automated lung sound classification using a hybrid CNN-LSTM network and focal loss function. Sensors, 22(3), 1232.
8. Roy, A., & Satija, U. (2023). Rdlinet: A novel lightweight inception network for respiratory disease classification using lung sounds. IEEE Transactions on Instrumentation and Measurement, 72, 1–13.
9. Sabry, A. H., Bashi, O. I. D., Ali, N. N., & Al Kubaisi, Y. M. (2024). Lung disease recognition methods using audio-based analysis with machine learning. Heliyon, 10(4).
10. Saky, S. M., Islam, M. R., Arefin, M. S., & Alam, S. (2025). Explainable multi-modal deep learning for automatic detection of lung diseases from respiratory audio signals. arXiv. https://arxiv.org/abs/2512.00563
11. Saky, S., & Tshering, U. (2025). Enhanced SegNet with integrated Grad-CAM for interpretable retinal layer segmentation in OCT images. arXiv. https://arxiv.org/abs/2509.07795
12. Sfayyih, A. H., Sulaiman, N., & Sabry, A. H. (2023). A review on lung disease recognition by acoustic signal analysis with deep learning networks. Journal of Big Data, 10(1), 101.
13. Shrivastava, P., Tripathi, N., Singh, B. K., et al. (2026). Extensive review: Speech analysis-based assessment of respiratory disorders applying machine learning and deep learning paradigms. Multimedia Tools and Applications, 85, 73. https://doi.org/10.1007/s11042-026-21202-z
14. Srivastava, A., Jain, S., Miranda, R., Patil, S., Pandya, S., & Kotecha, K. (2021). Deep learning based respiratory sound analysis for detection of chronic obstructive pulmonary disease. PeerJ Computer Science, 7, e369.
15. Vasava, R. P., & Joshiara, H. A. (2023). Different respiratory lung sounds prediction using deep learning. In 2023 4th International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 1626–1630). IEEE.
16. Wang, Q., Bu, Z., Mao, J., Zhu, W., Zhao, J., Du, W., Shi, G., Zhou, M., Chen, S., & Qu, J. (2024). Towards reliable respiratory disease diagnosis based on cough sounds and vision transformers. arXiv. https://arxiv.org/abs/2408.15667
17. World Health Organization. (2025, June 12). Chronic respiratory diseases: More than 80 million affected and many more undiagnosed, warns new WHO and European Respiratory Society report. https://www.who.int/europe/news/item/12-06-2025-chronic-respiratory-diseases--more-than-80-million-affected-and-many-more-undiagnosed--warns-new-who-and-european-respiratory-society-report
18. Xu, X., Zhang, Y., Niu, Q., Long, N., Li, J., Zhao, L., Yin, J., & Yang, J. (2026). Applications of speech analysis in diseases' assessment, prediction and diagnosis: A scoping review. Artificial Intelligence Surgery, 6, 114–149. https://doi.org/10.20517/ais.2025.70
19. Yadav, S., Agarwal, P., & Rizvi, S. A. M. (2024). Automated multiclass classification of lung diseases with deep learning: A hybrid CNN-LSTM approach. In 2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS) (pp. 1622–1627). IEEE.
20. Zhang, Y., Huang, Q., Sun, W., Chen, F., Lin, D., & Chen, F. (2024). Research on lung sound classification model based on dual-channel CNN-LSTM algorithm. Biomedical Signal Processing and Control, 94, 106257.