Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Detection of Voice and Lung Pathological Signal Using Acoustic Spectrogram Transformers

Revathi S.[1], Mohana Sundaram K.[2], Padmini Sharma[3], Manjusha Silas[4]

[1,2]Department of Electrical and Electronics Engineering, KPR Institute of Engineering and Technology, Coimbatore, India.
[3]Department of Electrical and Electronics Engineering, CSIT, Durg, Chhattisgarh, India.
[4]Department of Electrical Engineering, Christian College of Engineering and Technology, Bhilai, India.

E-mail: [1]revaviji23@gmail.com, [2]kumohanasundaram@gmail.com, [3]padminisharma@csitdurg.in, [4]m.silas@ccetbhilai.ac.in

Abstract

In the medical field, identifying various pathological conditions poses a crucial challenge because it requires an invasive and contact-based data extraction technique. Therefore, non-invasive and non-contact forms of vital data, such as speech signals, can be used to identify various pathological conditions. Speech signals have distinguishing phonetic characteristics that change when a pathological condition occurs in the human body. By using these changes, various pathological signals can be classified by training machine learning and deep learning models with the acoustic features of speech signals. This work proposes the acoustic spectrogram transformer, where all the layers in the transformer are trained using acoustic characteristics extracted from the speech signals of voice and lung disease patients. Mel-frequency cepstral coefficients (MFCCs), Mel spectrograms, and spectral variables like centroid, bandwidth, roll-off, and zero-crossing rate are used for feature extraction from the voice and lung dataset. These acoustic features train the transformer blocks and depth-adaptive parameters, enabling the model to capture complex patterns for effective signal classification. Along with this architecture, the model consists of frequency-focused attention mechanisms used to extract spectral characteristics that are most indicative of pathological conditions. Meanwhile, multiple pooling strategies are employed for the effective aggregation of temporal information. Due to this targeted design, the system serves as an effective clinical tool for classification, minimizing computational complexity and achieving an accuracy of about 83% in voice pathology classification and 99% in lung pathology classification.

Keywords: Voice Pathology, Lung Pathology, Acoustic Spectrogram Transformer, Mel Spectrogram.

1. Introduction

Voice disorders affect a significant section of the world's inhabitants, with rates commonly reported to be between 3-9% among general populations. They are more prevalent in verbally strenuous jobs such as teaching [1], singing, and call center operations. Similarly,

DOI: https://doi.org/10.36548/jiip.2025.4.012
Received: 17.09.2025, received in revised form: 15.10.2025, accepted: 28.10.2025, published: 07.11.2025

lung diseases affect individuals by causing difficulties during breathing and reducing the oxygen supply to the body, which may lead to premature death. These disorders greatly impact the quality of life, professional capabilities, and emotional well-being of individuals affected by voice and lung diseases.

Voice and lung pathologies are often diagnosed through invasive examinations such as laryngoscopy or through perceptual methods involving subjectivity from clinicians [2]. For lung diseases, CT scans or ultrasounds are required. This diagnostic process is prone to variability and often overlooks the acoustic features essential for characterizing the early stages of disorders in voice and lung signals [3] [4]. The introduction of deep learning techniques [5] has revolutionized signal processing in various applications such as speech and audio analysis for early detection and explicit classification of voice and lung pathological signals in non-invasive diagnostic tools [6]. Recently, voice pathological signals can be detected and classified using transformer-based architectures, which have shown impressive results in computer vision and natural language processing fields [7]. Through their self-attention processes, transformers, such as the Vision Transformer [8] [9], outperform conventional convolutional or recurrent neural network architectures in modeling local as well as global dependencies in sequential data.

For voice and lung pathological classification, an Audio Spectrogram Transformer (AST) is proposed, with hyperparameters specially optimized to fit the model for pathological classification. The proposed model distinguishes signals from normal voices and pathological voices, such as dysphonia, laryngitis, laryngeal nerve palsy, bronchiectasis, COPD, and pneumonia. These illnesses pose a challenging multi-class prediction problem since they represent a wide range of vocal and lung abnormalities with unique changes in acoustic characteristics. The dataset used in this study consists of carefully chosen audio recordings from healthy subjects and their sick counterparts bearing those voice and lung abnormalities, providing a strong foundation for model development and evaluation [10] [11]. This work innovates in different critical areas of voice and lung pathology classification.

To overcome the imbalance in voice and lung pathological data, an augmentation approach is used, which includes adding calibrated noise, pitch shifting, time stretching, and innovative sample mixing [12] [13] strategies that preserve the essential pathological characteristics while increasing training data diversity. Furthermore, a class-aware balancing strategy is employed that generates synthetic samples through targeted augmentation to ensure equivalent representation across all diagnostic categories, preventing the model from developing bias toward more prevalent conditions. Many feature extraction techniques have been used in early studies, such as Wav2Vec feature extraction, which utilizes a transformer technique to effectively extract acoustic features by converting signals into spectrogram transformers [14]. A pre-trained transformer model has been used to detect abnormal signals in patients with cleft lips [15]. Many transformer models have been developed for various pathological detections. This work proposes audio spectrogram transformers that classify voice and lung pathological signals from healthy signals using speech datasets. The complete workflow from raw signal to diagnosis is illustrated in Figure 3. The contributions of this research can be highlighted as follows:

- To develop a dataset that is collected from various publicly available sources, which are gathered from the affected individuals.

- To extract the frequency characteristics from the acoustic input signal using the Mel-Frequency Cepstral Coefficients (MFCCs), Mel Spectrogram, and spectral variables like centroid, bandwidth, roll-off, and zero crossing rate.
- To design the audio spectrogram transformer to classify various pathological signals from normal signals in the most efficient and effective form.
- To evaluate the performance of the audio spectrogram transformer for multi-class classification tasks.

The research paper is organized as follows: Section 2 discusses the recent work carried out in identifying voice and lung pathological conditions; Section 3 presents the dataset used, feature extraction techniques, and proposed audio spectrogram transformer model; Section 4 shows the performance of the model in classifying voice and lung signals; and finally, Section 5 concludes the work performed.

2. Related Work

Audio spectrogram transformers are used for the classification of voice disorders by applying a transformer-based deep learning architecture. The multi-layer feature fusion and pooling combination is employed for sound classification by modifying the self-supervised audio spectrogram transformer [16]. A hybrid LSTM-transformer model is used for emotion recognition from audio speech signals [17]. For the classification of neurological disorders, vision transformers are utilized to classify human speech [18]. A transformer model combining convolution and transformation is used for heart sound classification to detect cardiovascular diseases [19]. Crying samples of infants are used to enhance pediatric healthcare using transformer-based approaches to diagnose pathological signals [7]. Lung sounds can be classified using audio spectrogram transformers by adaptively modifying the transformer model [20]. Swin transformers are employed for classifying dysarthria by capturing local features from the voice signal [21]. Voice pathological signals can be classified using optimized convolutional neural networks [22] and fast learning networks [23] by extracting MFCC and linear predictive coding (LPC) from real-time signals.

3. Materials and Methodology

3.1 Dataset Description

3.1.1 Voice Dataset

SVD is a medical speech database that was developed in Saarbrücken, Germany, by the Institute of Phonetics at Saarland University [10]. It includes voice recordings from 2,000 participants. These recordings were in WAV format and recorded with a sampling frequency of 50 kHz. Participants include healthy speakers and patients with various speech disorders. Data were gathered through a dual-channel recording: the electroglottographic (EGG) signal was recorded simultaneously using surface electrodes placed bilaterally around the participant's neck at the thyroid cartilage level, and the acoustic signal was recorded through a high-quality condenser microphone placed at a standard distance of 30 cm from the speaker's mouth. To maintain acoustic consistency, all recordings were made in a sound-attenuated booth with a background noise level of less than 40 dB. The recordings consist of running speech samples,

standardized text passages, and sustained vowels—exceptionally a, i, and u—at high, low, low-high, and normal pitches. The acoustic signal includes recordings from both voice pathologies and healthy patients at a size of 194 KB in .wav format. The dataset comprises participants aged 18-75 years, with a distribution of 55% female and 45% male. For this study, 4,383 recordings from healthy subjects and 3,647 recordings from pathological conditions were utilized. The pathological cases included 700 dysphonias, 900 laryngitis cases, and 2,047 cases of laryngeal nerve palsy. The classification of voice pathology was performed by two experienced speech-language pathologists with more than 5 years of clinical experience, and all recordings were evaluated using Grade, Roughness, Breathiness, Asthenia, and Strain (GRBAS) annotators.

3.1.2 Respiratory Dataset

The respiratory sound database consists of recordings of breathing sounds and respiratory audio data used for pulmonary disorder classification [11]. It was created at the 2017 Respiratory Sound Database International Conference on Biomedical Health Informatics (ICBI). The database contains 920 annotated recordings obtained from 126 subjects, both healthy and with a wide variety of respiratory conditions. Recordings were made using three standardized devices: (1) AKG C417L miniature condenser microphone for tracheal sounds, (2) 3M Littmann Classic II SE stethoscope for chest auscultation, and (3) WelchAllyn Meditron Master Elite Electronic Stethoscope for digital sound capture. Recordings were taken at each participant's anterior, posterior, and lateral chest position, following a standardized protocol and taking 10-20 seconds per recording position. Single-channel and multi-channel recording configurations were employed based on the research objective. The acoustic signals are stored in WAV format with an average file size of 2.52 MB per recording, sampled at 44.1 kHz with 16-bit resolution. The acoustic signal is a recording from voice pathologies and healthy patients at the 2.52 MB size in .wav format. The cohort includes participants aged 6-90 years, with a distribution of 58% male and 42% female. For pulmonary disorder classification, this study utilized 35 recordings from healthy participants and 846 recordings from patients with respiratory pathologies, including 16 cases of bronchiectasis, 793 cases of chronic obstructive pulmonary disease (COPD), and 37 cases of pneumonia.

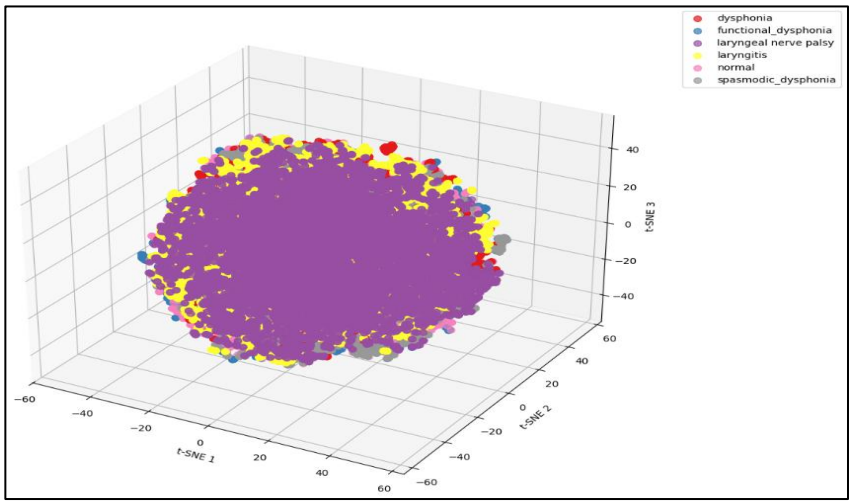


Figure 1. 3D t-SNE of Voice Dataset Distribution After Augmentation

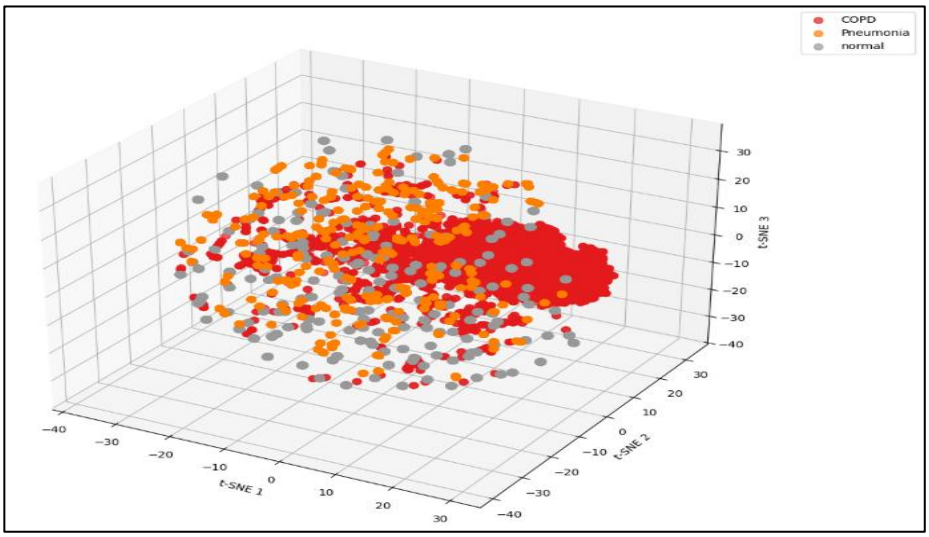


Figure 2. 3D t-SNE of Lung Dataset Distribution After Augmentation

To balance the dataset the augmentation technique such as adding noise, changing pitch, and stretching was employed [12] [13]. Gaussian noise with a small variance of 0.01 was used for noise addition in augmented data generation. The change pitch technique was used to generate the augmented signal were achieved by amplitude scaling of 2 while preserving timing without altering the vocal characteristics and lung capacity. Changes in temporal characteristics by means of speeding up or slowing down at a rate of 0.8 without affecting pitch, generated using stretching augmentation techniques. The dataset distribution across the classes after applying the augmentation technique is illustrated in Figures 1 and 2 in 3D t-SNE (t-Distributed Stochastic Neighbor Embedding).

3.2 Feature Extraction

In this work, to extract comprehensive acoustic information from audio files, classical parameters are used: Mel-Frequency Cepstral Coefficients, Mel Spectrogram, and spectral variables such as centroid, bandwidth, roll-off, and zero crossing rate [24]. The mel-frequency scale has particular use in the speech signal analysis domain due to its mimicry of the human auditory system, providing a reduced resolution at higher frequencies and improved frequency resolution at lower frequencies. Feature extraction begins with preprocessing to ensure regularized and robust analysis. First, audio signals are loaded with a sampling rate of 22.05 kHz, which is a good trade-off between frequency resolution for speech analysis and computational efficiency. Noise reduction is achieved through amplitude thresholding, where signals with an amplitude of less than 0.005 are set to zero, thus removing low-level background noise. Audio length normalization ensures that all recordings are at least one second long, which is sufficient for speech analysis while keeping it under ten seconds for relevant speech segments.

The principal feature set includes 26-dimensional MFCC [25] [26] vectors extracted with a frame size of 2007 samples to preserve temporal dynamics that are important for pathological classification. Alongside MFCCs, mel-spectrogram features are extracted using 80 mel-frequency bands spanning 80 Hz to 8000 Hz, corresponding to the main speech frequency domain. The mel-spectrogram is converted to a decibel scale so that it correlates to logarithmic human auditory perception. Spectral features such as spectral centroid, bandwidth

and roll off are extracted to characterize voice quality aspects directly related to pathological conditions. Some temporal features including the zero-crossing rate (ZCR) measure the rate at which an audio signal crosses zero amplitude [27]. Features have all been extracted consistently with uniform parameters across all dataset members, impacting reproducibility and comparability. While established MFCC coefficients are considered proven and reliable in their own right, the extra spectral and temporal features would further cement pathology detection. In MFCC, first and second order temporal derivatives are used to capture dynamic information to avoid the redundancy of higher order statistics and it uses 26 MFCC coefficients and 80 mel bands for noise reduction. The mel spectrogram uses the speech computation range of 80 – 8000Hz to eliminate frequency redundancy. In cases of spectral variable non overlapping spectral extraction, methods are used to eliminate redundant frequencies from multiple correlated spectral descriptors.

3.3 Acoustic Spectrogram Transformer

This Acoustic Spectrogram Transformer [7] is a highly sophisticated transformer-based architecture designed for voice and lung pathology detection, processing audio features and classifying normal voices and pathological ones such as dysphonia, laryngitis, laryngeal nerve palsy, Bronchiectasis, COPD and Pneumonia. The model begins with dual-pathway feature embedding where two parallel dense networks with different complexities process the input, followed by a learnable gating mechanism that intelligently combines both pathways to create optimal initial representations. Learnable positional encodings are added to provide temporal sequence information, crucial for audio analysis, followed by layer normalization for training stability. The core architecture consists of 6 transformer encoder blocks with progressive complexity scaling, where deeper layers have more attention heads, larger dimensions, and higher dropout rates to capture increasingly complex patterns as shown in Figure 3.

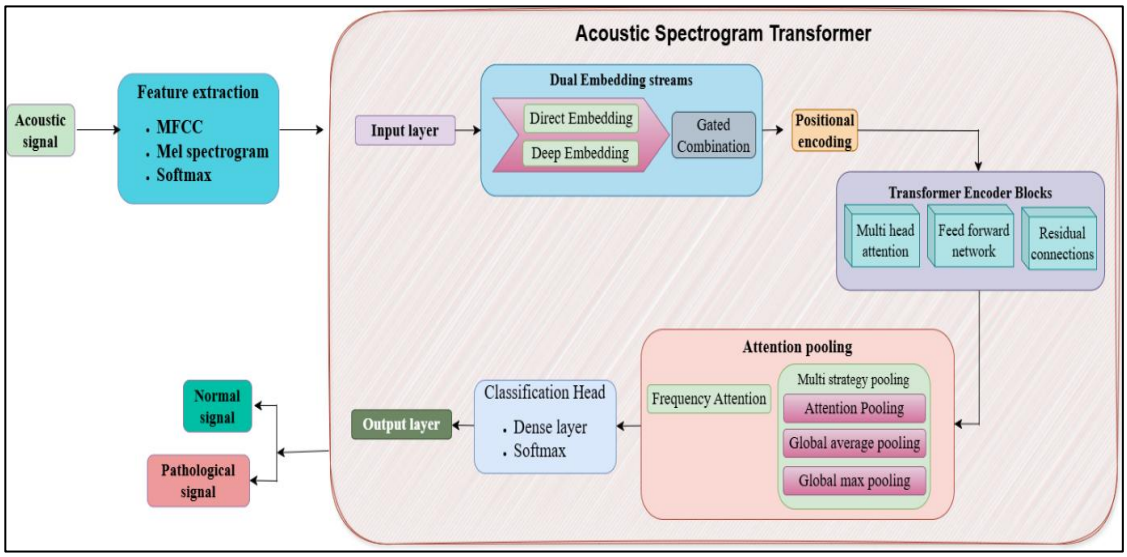


Figure 3. Workflow from Raw Signal to Diagnosis Using Acoustic Spectrogram Transformer Architecture

Each transformer block uses pre-layer normalization, 8-head multi-head attention with 64-dimensional keys, extensive batch normalization, and a two-stage feed-forward network that expands from 128 to 256 dimensions before projecting back. Residual connections occur both within blocks and between every two blocks to maintain gradient flow and enable learning of

local and global patterns. After transformer processing, the model employs sophisticated feature aggregation through frequency attention that learns importance weights from different sequence parts, combined with three pooling strategies: attention-weighted sum, global average and max pooling. In these multi-pooling strategies, the attention weighted sum focuses on salient frequency bands to capture the pathological information, global average pooling provides robust overall acoustic characteristics resistant to noise and global max pooling captures peak anomalies like voice breaks and tremors in pitch. A frequency focused attention mechanism is used to incorporate the frequency domain to operate in the spatial or temporal domain and Fourier transformer is used to capture patterns and global dependencies in the frequency domain. The multi-head attention mechanism allows the model to simultaneously focus on multiple acoustic aspects such as pitch, energy and timbre at different time scales, which is essential for distinguishing between normal and pathological voice. Additionally, the transformer's global context modelling capability outperforms CNNs limited receptive fields and RNNs' sequential processing bottlenecks, while the integrated frequency attention mechanism and multiple pooling strategies provide more sophisticated feature extraction specifically suited for pathology detection. A squeeze and excitation block further enhances features by learning channel-wise attention weights, reducing dimensionality before expanding back with sigmoid gating. The classification head uses progressive dimensionality reduction through three dense layers with extensive batch normalization, Swish activation, and dropout, culminating in a softmax output layer.

Table 1. Various Parameters of Audio Spectrogram Transformer

Component	Parameter	Value
Embedding system – stream 1	Dense units	128
	Key dimension	64
	L2 Regularization	0.0001
Embedding system – stream 2	First dense layer units	64
	Second dense layer units	128
	Activation	LeakyReLU
Dropout	Spatial dropout rate	0.2
Transformer backbone	Number of encoder blocks	6
	Attention heads (early layers)	8
	Attention heads (deep layers)	11
	Key dimension (early layers)	64
	Key dimension (deep layers)	85
	Feed-forward dimensions (early)	256

	Feed-forward dimensions (deep)	384
	Dropout rate	0.3
Classification head	First layer units	256
	Second layer units	128
	Third layer units	64
	L2 Regularization	0.0002
Optimizer (Adam)	Learning rate	0.0001
	Beta 1	0.9
	Beta 2	0.999
	Gradient Clipping Norm	1.0

The classification head implements a bottleneck architecture, reducing dimensions from 256 to 128 to 64 units before the final softmax layer, with L2 regularization increasing to 0.0002 in these critical layers. The Adam optimizer operates with a learning rate of 0.00025, beta values of 0.9 and 0.999, and implements gradient clipping at a norm of 1.0 to prevent exploding gradients as shown in Table 1. The parameters used in this model are selected using an ablation study, where the model shows less overfitting when adding layer normalization and L2 regularization. The model achieves 78% accuracy with attention heads in the early layer and by adding the deep attention head layer, the model shows higher accuracy.

Table 2. Optimization Parameter of Voice and Lung Model

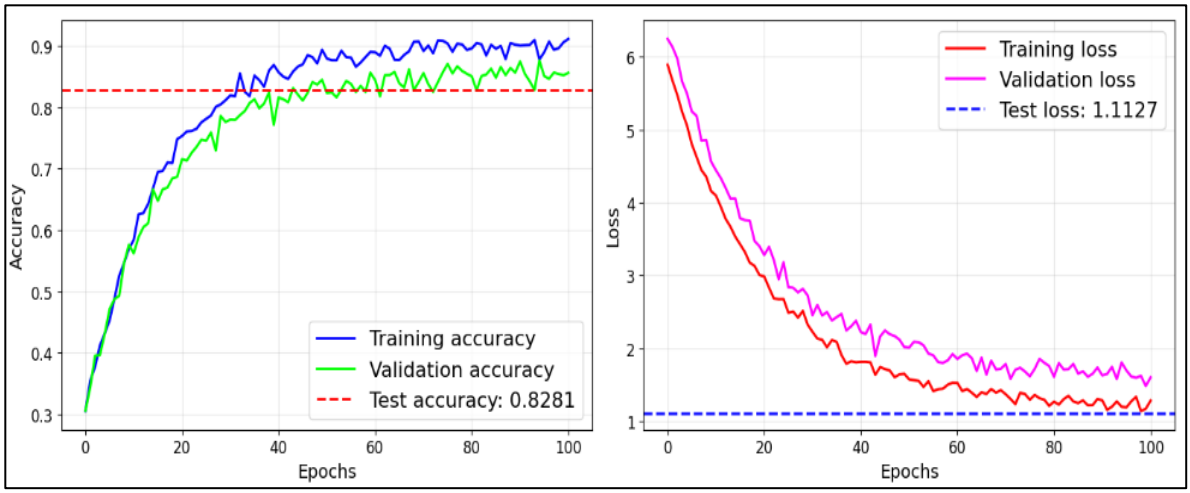
Training configuration	**Grid search range**	**Value**
Multi head attention	4 to 16	8
Dropout rate	0.1 to 0.5	0.3
Optimizer	Adam, AdamW and RMSprop	AdamW
Learning rate	0.001-0.0001	0.0001
Batch size	16 and 32	32

The algorithm is optimized using 8 multi head attention, a dropout rate of 0.3, AdamW with a learning rate 0.0001, gradient clipping, early stopping, and learning rate of scheduling, specifically designed to handle the temporal nature of audio features while maintaining both regional temporal patterns and global voice characteristics essential for precise pathology detection as shown in Table 2. The hyperparameters of the acoustic spectrogram transformer are optimally selected using the high test accuracy and minimal test loss. Training was conducted on an Intel Core i7-13700 system, with the entire evaluation process taking

approximately 0.54 seconds for testing. This architecture combines transformer attention mechanisms with domain-specific optimizations for robust voice analysis, utilizing comprehensive regularization strategies including batch normalization, progressive dropout, and L2 regularization to ensure stable training and prevent overfitting in medical classification tasks.

4. Result and discussion

The proposed transformer model is used to classify normal and pathological signals in various multi-pathological classifications. During the training phase of the transformer model, the voice model achieves 83% accuracy and shows less overfitting. Figure 4a shows the training and validation accuracy of the acoustic spectrogram transformer. Validation and training accuracy rise with increasing epochs until they reach the maximum epochs, and Figure 4b illustrates the statistical decline in training and validation loss over the epochs.



(a) (b)

Figure 4. a) Training and Validation Accuracy Curve b) Training and Validation Loss Curve

Table 3. Performance Analysis of Voice Pathological Signal

Voice signal	Precision	Recall	F1-Score
Laryngeal nerve palsy	0.83	0.77	0.80
Laryngitis	0.86	0.89	0.88
Dysphonia	0.82	0.95	0.88
Normal	0.80	0.70	0.75

Table 3 illustrates the model's performance by displaying the acoustic spectrogram transformer's precision, recall, and F1-score for the vocal pathological categorization of normal to pathological signals, which include dysphonia, laryngeal nerve palsy, and laryngitis. The model shows about 83% accuracy during the training and testing phases of the acoustic

spectrogram transformer. Figure 5 represents the confusion matrix, which demonstrates the voice classification model's performance across four distinct categories: laryngeal nerve palsy, laryngitis, dysphonia, and normal voice conditions.

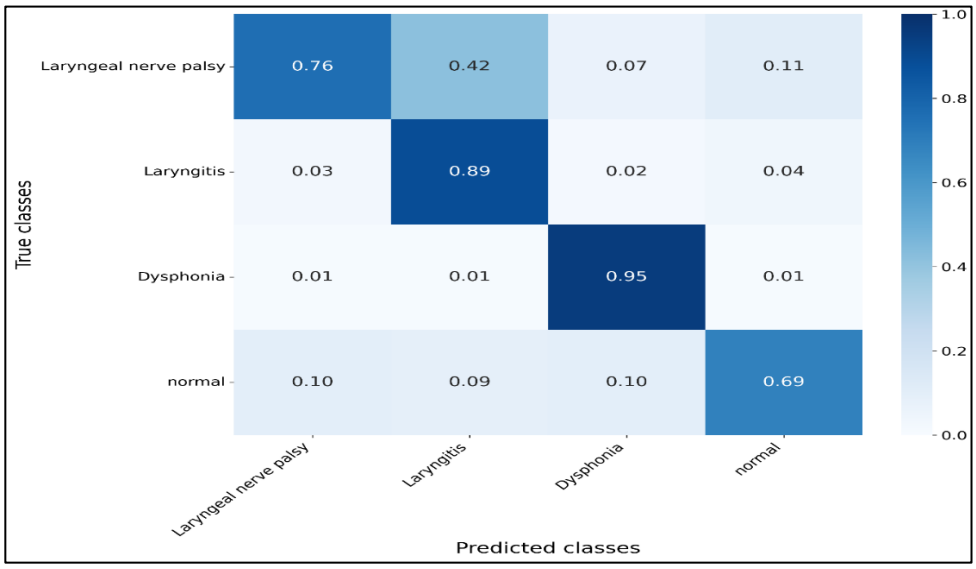
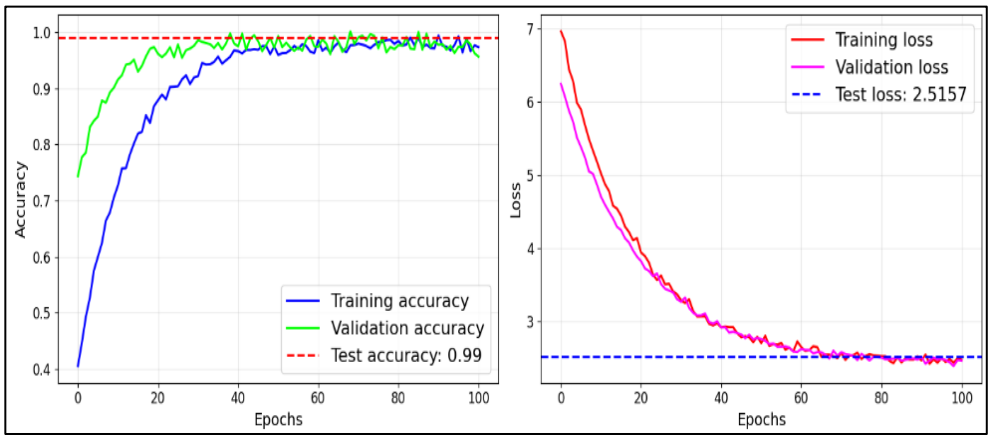


Figure 5. Confusion Matrix for Vocal Signal Classification

The model exhibits strong overall performance with high diagonal values, indicating effective discrimination between different voice pathologies. The model specifically shows remarkable classification accuracy rates for pathological diseases, with dysphonia exhibiting the best classification accuracy at 95%, closely followed by laryngitis at 89%, and laryngeal nerve palsy at 76%. However, the classification of normal voices presents a notable challenge, with only 69% accuracy in this category. The acoustic spectrogram transformer will work on classifying the respiratory diseases bronchiectasis, COPD and pneumonia by using wheezes and crackles occurring during the respiratory cycle. Figure 6 a) depicts the training and validation accuracy of the model on the lung signal classification. The training and validation accuracy curve doesn't show any deviation and it increases over the number of epochs. Figure 6 b) shows the loss curve for the model in decreasing order.



(a) (b)

Figure 6. a) Training and Validation Accuracy for Lung Signal Classification b) Training and Validation Loss Curve for Lung Signal Classification

Table 4. Performance Analysis of Lung Pathological Signal

Lung signal	Precision	Recall	F1-Score
Bronchiectasis	0.98	1.00	0.99
COPD	1.00	0.94	0.97
Pneumonia	0.97	1.00	0.98
Healthy	1.00	1.00	1.00

Table 4 illustrates the model's performance by displaying the acoustic spectrogram transformer's precision, recall, and F1-score for the lung pathological categorization of normal to pathological signals, which include Bronchiectasis, COPD, Pneumonia, and Healthy. The model shows 99% accuracy in classifying lung signals for respiratory disease identification, as illustrated in Figure 7. The confusion matrix presents the performance analysis of lung diseases in classifying normal and various pathological signals such as Bronchiectasis, COPD, and Pneumonia. The model obtained a 95% confidence interval for the given input to both the vocal and lung datasets for the acoustic spectrogram transformer model, as given in Table 5, and the p-value is equal to zero upon testing the model for statistical validation.

Table 5. Metric Comparison of Confidence Intervals for Vocal and Lung Model

	Vocal model		Lung model	
Metric	Mean	Confidence Intervals	Mean	Confidence Intervals
Accuracy	0.824	[0.811, 0.836]	0.987	[0.986, 0.988]
Precision	0.823	[0.810, 0.836]	0.989	[0.988, 0.990]
Recall	0.824	[0.812, 0.836]	0.990	[0.986, 0.992]
F1	0.821	[0.808, 0.833]	0.988	[0.986, 0.988]

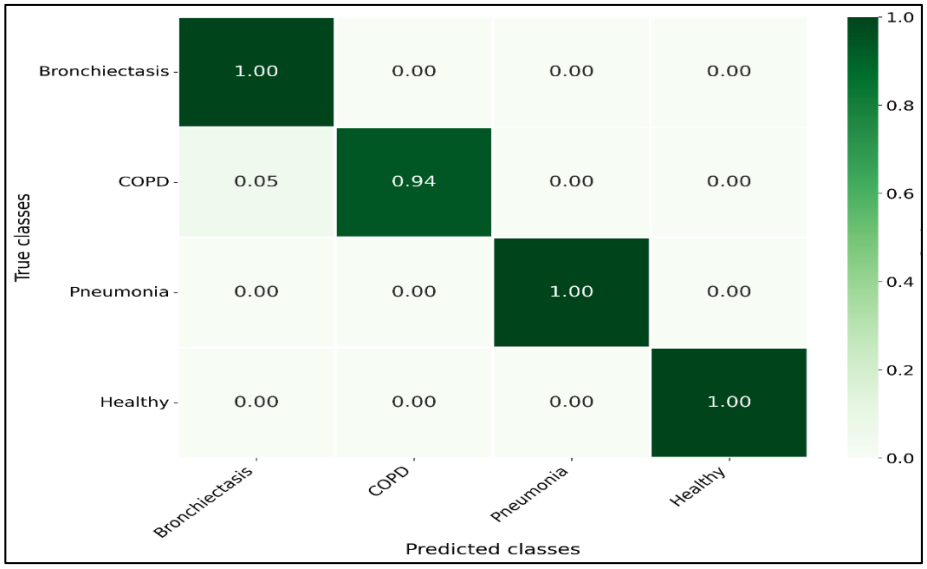


Figure 7. Confusion Matrix for Lung Signal Classification

5. Discussion

The developed acoustic spectrogram transformer is used to classify voice and lung pathological conditions from speech data. The model shows the best results in classifying the pathological conditions. For the voice dataset, the model yields 83% accuracy, and it achieves 99% accuracy for lung signal classification. Various transformers and deep learning models, such as the audio spectrogram transformer [7], vision transformer [18], swim transformer [21], and convolutional neural network [22], are used to compare the results; among them, the proposed model shows the highest performance in categorizing the pathological signals, as shown in Table 6. The proposed acoustic spectrogram transformer demonstrates close tracking between training and validation accuracy, which indicates that the model does not show any signs of overfitting.

Table 6. Comparative Analysis of Proposed Work with the Existing Transformer Model for the Voice and Lung Dataset

Ref	Model	Voice model				Lung model			
			Precision	Recall	F1-Score		Precision	Recall	F1-Score
[7]	Audio spectrogram transformer	Laryngeal nerve palsy	0.31	0.30	0.40	Bronchiectasis	0.57	0.58	0.53
		Laryngitis	0.29	0.31	0.43	COPD	0.50	0.59	0.55
		Dysphonia	0.30	0.32	0.47	Pneumonia	0.57	0.60	0.58
		normal	0.32	0.29	0.40	Healthy	0.55	0.50	0.51
[18]	Vision transformer	Laryngeal nerve palsy	0.80	0.79	0.85	Bronchiectasis	0.99	0.99	0.97
		Laryngitis	0.83	0.80	0.80	COPD	0.98	0.96	1.00
		Dysphonia	0.79	0.84	0.87	Pneumonia	0.95	0.96	0.95
		normal	0.80	0.83	0.80	Healthy	1.00	0.99	0.99
[21]	Swin transformer	Laryngeal nerve palsy	0.74	0.72	0.71	Bronchiectasis	0.85	0.84	0.86
		Laryngitis	0.73	0.75	0.75	COPD	0.90	0.89	0.91
		Dysphonia	0.69	0.67	0.68	Pneumonia	0.89	0.91	0.88
		normal	0.70	0.73	0.73	Healthy	0.89	0.9	0.91

[22]	Convolutional neural network	Laryngeal nerve palsy	0.78	0.79	0.78	Bronchiectasis	0.95	0.97	0.97
		Laryngitis	0.81	0.80	0.80	COPD	0.91	0.94	1.00
		Dysphonia	0.78	0.82	0.80	Pneumonia	0.90	0.97	0.93
		normal	0.79	0.79	0.79	Healthy	0.98	0.99	0.97
Proposed Work		Laryngeal nerve palsy	0.83	0.77	0.80	Bronchiectasis	0.98	1.00	0.99
		Laryngitis	0.86	0.89	0.88	COPD	1.00	0.94	0.97
		Dysphonia	0.82	0.95	0.88	Pneumonia	0.97	1.00	0.98
		normal	0.80	0.70	0.75	Healthy	1.00	1.00	1.00

6. Conclusion

This paper presents the performance of an Acoustic Spectrogram Transformer in the multi-class classification of pathologies and achieves very impressive results for voice and lung pathology detection. The contribution of the work relies mainly on the architectural design and deep feature extraction strategy. Mel-frequency cepstral coefficients and Mel spectrograms are successfully leveraged to capture temporal and spectral features relevant to disease classification tasks. The effectiveness of our strategy is supported by experimental results that show impressive accuracy rates of 99% for lung pathology classification and roughly 83% for voice pathological classification. These metrics demonstrate the model's clinical viability and suitability for medical applications. The high accuracy obtained for the task of lung pathology detection, in particular, underlines the versatility of the AST framework for solving various acoustic pathology detection challenges. The proposed framework could be extended to other pathology conditions, and its cross-lingual performance for a variety of patient populations may be studied. Additionally, it can be integrated into real-time clinical decision support systems.

Declarations

Ethics Approval and Consent to Participate

Not Applicable.

Consent for Publication

Not Applicable

Funding

Not Applicable

Availability of Data and Materials

Data will be available based on the request

Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author's Contributions

RS, KM, PS and MS prepared the manuscript, RS drawn figures. All authors reviewed the manuscript.

Acknowledgements

Not Applicable

References

[1] Sankar, G., V. Ganesan, R. V. Shantharam, K. Palanisamy, and I. Katam. "Epidemiology of Voice Disorders Among Government School Teachers-An Analytical Cross-Sectional Study from Kancheepuram District." (2022). National Journal of Community Medicine 13 (2023): 869–875.

[2] Abdulmajeed, Nuha Qais, Belal Al-Khateeb, and Mazin Abed Mohammed. "A Review on Voice Pathology: Taxonomy, Diagnosis, Medical Procedures and Detection Techniques, Open Challenges, Limitations, and Recommendations for Future Directions." Journal of Intelligent Systems 31 (2022): 855–875.

[3] Fujiki, Robert Brinton, and Susan L. Thibeault. "Voice Disorder Prevalence and Vocal Health Characteristics in Children." JAMA Otolaryngology–Head & Neck Surgery 150, no. 8 (2024): 677-687.

[4] Kaliappan, Vishnu Kumar, Rajasekaran Thangaraj, P. Pandiyan, K. Mohanasundaram, S. Anandamurugan, and Dugki Min. "Real-Time Face Mask Position Recognition System Using YOLO Models for Preventing COVID-19 Disease Spread in Public Places." International Journal of Ad Hoc and Ubiquitous Computing 42, no. 2 (2023): 73-82.

[5] Revathi, S., and K. Mohana Sundaram. "Deep Learning-Based Voice Pathology Detection from Electroglottography." In Approaches to Human-Centered AI in Healthcare, IGI Global Scientific Publishing, 2024, 236-257.

[6] Ramalingam, Anbukarasi, and Nithya Narayanan. "The Diagnostic Efficacy of Flexible Fiberoptic Laryngoscopy and Its Correlation with Histopathology in Different Benign Lesions of the Vocal Cord in a Tertiary Care Hospital: A Prospective Clinical Study." Cureus 16, no. 12 (2024).

[7] Tami, Mohammad, Sari Masri, Ahmad Hasasneh, and Chakib Tadj. "Transformer-based Approach to Pathology Diagnosis Using Audio Spectrogram." Information 15, no. 5 (2024): 253.

[8] Hegde, K. Jayashree, K. Manjula Shenoy, and K. Devaraja. "Performance Evaluation of Pre-Trained Models for Classification of Vocal Cord Paralysis over Vowels." Paper presented at the Second International Conference on Networks, Multimedia and Information Technology (NMITCON), Bengaluru, India, 2024.

[9] Hemmerling, Daria, Marek Wodzinski, Juan Rafael Orozco-Arroyave, David Sztaho, Mateusz Daniol, Pawel Jemiolo, and Magdalena Wojcik-Pedziwiatr. "Vision Transformer for Parkinson's Disease Classification Using Multilingual Sustained Vowel Recordings." In 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), IEEE, 2023, 1-4.

[10] Woldert-Jokisz, B. "Saarbruecken Voice Database, Version 2.0." 2007. https://stimmdb.coli.unisaarland.de/index.php4#target.

[11] "Respiratory Sound Database, Version 2." Kaggle. https://www.kaggle.com/datasets/vbookshelf/respiratory-sound-database.

[12] Javanmardi, Farhad, Sudarsana Reddy Kadiri, and Paavo Alku. "A Comparison of Data Augmentation Methods in Voice Pathology Detection." Computer Speech & Language 83 (2024): 101552.

[13] Shen, Jiakun, Xueshuai Zhang, Yu Lu, Pengfei Ye, Pengyuan Zhang, and Yonghong Yan. "Novel Audio Characteristic-Dependent Feature Extraction and Data Augmentation Methods for Cough-Based Respiratory Disease Classification." Computers in Biology and Medicine 179 (2024): 108843.

[14] Cai, Jie, Yuliang Song, Jianghao Wu, and Xiong Chen. "Voice Disorder Classification Using Wav2vec 2.0 Feature Extraction." Journal of Voice (2024). 0892-1997.

[15] Bhattacharjee, Susmita, Hanumant Singh Shekhawat, and S. R. M. Prasanna. "Classification of Cleft Lip and Palate Speech Using Fine-Tuned Transformer Pretrained Models." In International Conference on Intelligent Human Computer Interaction, Cham: Springer Nature Switzerland, 2023, 55-61.

[16] Choi, Hyosun, Li Zhang, and Chris Watkins. "Dual Representations: A Novel Variant of Self-Supervised Audio Spectrogram Transformer with Multi-Layer Feature Fusion and Pooling Combinations for Sound Classification." Neurocomputing 623 (2025): 129415.

[17] Andayani, Felicia, Lau Bee Theng, Mark Teekit Tsun, and Caslon Chua. "Hybrid LSTM-Transformer Model for Emotion Recognition from Speech Audio Files." IEEE Access 10 (2022): 36018-36027.

[18] Soylu, Emel, Sema Gül, Kübra Aslan, Muammer Türkoğlu, and Murat Terzi. "Vision Transformer Based Classification of Neurological Disorders from Human Speech." Firat University Journal of Experimental and Computational Engineering 3, no. 2 (2023): 160-174.

[19] Cheng, Jiawen, and Kexue Sun. 2023. "Heart Sound Classification Network Based on Convolution and Transformer" Sensors 23, no. 19: 8168.

[20] Xiao, Li, Lucheng Fang, Yuhong Yang, and Weiping Tu. "LungAdapter: Efficient Adapting Audio Spectrogram Transformer for Lung Sound Classification." In Proc. Interspeech 2024, 2024, 4738-4742.

[21] Mahum, Rabbia, Ismaila Ganiyu, Lotfi Hidri, Ahmed M. El-Sherbeeny, and Haseeb Hassan. "A novel Swin transformer based Framework for Speech Recognition for Dysarthria." Scientific Reports 15, no. 1 (2025): 20070.

[22] Farazi, Sahar, and Yaser Shekofteh. "Efficient DL Models for Voice Pathology Detection in Healthcare Applications using Sustained Vowels." Journal of Innovations in Computer Science and Engineering (JICSE) 2, no. Special Issues 2 (2025): 26-32.

[23] Albadr, Musatafa Abbas Abbood, Masri Ayob, Sabrina Tiun, Fahad Taha AL-Dhief, Muataz Salam Al-Daweri, Raad Z. Homod, and Ali Hashim Abbas. "Fast Learning Network Algorithm for Voice Pathology Detection and Classification." Multimedia Tools and Applications 84, no. 17 (2025): 18567-18598.

[24] Kumar, Deepak, Udit Satija, and Preetam Kumar. "Analysis and Classification of Electroglottography Signals for the Detection of Speech Disorders." In 2023 National Conference on Communications (NCC), IEEE, 2023, 1-6.

[25] Rao, PVL Narasimha, and S. Meher. "ORG-RGRU: An Automated Diagnosed Model for Multiple Diseases by Heuristically based Optimized Deep Learning using Speech/Voice Signal." Biomedical Signal Processing and Control 88 (2024): 105493.

[26] Devi, Kharibam Jilenkumari, Ayekpam Alice Devi, and Khelchandra Thongam. "Automatic Speaker Recognition using MFCC and Artificial Neural Network." Int. J. Innov. Technol. Explor. Eng 9, no. 1 (2019): 39-42.

[27] Wu, Yuanbo, Changwei Zhou, Ziqi Fan, Di Wu, Xiaojun Zhang, and Zhi Tao. "Investigation and Evaluation of Glottal Flow Waveform for Voice Pathology Detection." IEEE Access 9 (2020): 30-44.