STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **14**, November 2025, pp 0–21.
Published online in International Academic Press (www.IAPress.org)

Cross-Attention Transformer Networks with Optimized Feature Selection for Explainable Respiratory Disease Classification

Bajeszeyadaljunaeidia[2,*], Mohammed Tawfik[1,*], Issa M. Alsmadi[2], Yasser Mohammad Al-Sharo[1]

[1]*Department of Cyber Security, Faculty of Information Technology, Ajloun National University, P.O.43, Ajloun-26810, Jordan*
[2]*Faculty of Information and Technology, Ajloun National University, P.O. Box 43, Ajloun 26810, Jordan*

Abstract Respiratory diseases require accurate diagnosis for effective treatment, yet traditional methods rely on subjective assessments and expensive procedures. This paper presents a transformer-based cross-attention framework for acoustic respiratory disease classification with explainable AI integration. The CrossAttentionAcousticNetwork combines CNN spectral feature extraction with transformer temporal modeling, enhanced by cross-attention mechanisms for multi-modal feature fusion. Namib Beetle Optimization selects discriminative features from 100-dimensional handcrafted and deep spectral representations, while LIME provides clinical interpretability. Evaluation on ICBHI 2017 and KAU datasets achieves 99.0% and 95.0% accuracy respectively, representing 21.39% improvement over existing methods. The framework demonstrates superior performance across asthma, COPD, pneumonia, and heart failure classifications while maintaining computational efficiency for real-time deployment. Integrated explainable AI reveals clinically relevant acoustic patterns, with data augmentation improving minority class recognition by 19%. This approach bridges the gap between high-performance deep learning and clinical transparency requirements for automated respiratory disease screening.

Keywords Respiratory classification, transformer networks, cross-attention, explainable AI, feature optimization, clinical systems

DOI: 10.19139/soic-2310-5070-2912

1. Introduction

Respiratory diseases represent a significant global health burden, affecting millions of individuals worldwide and contributing substantially to morbidity and mortality rates. Traditional diagnostic approaches rely heavily on subjective clinical assessments, chest imaging, and invasive procedures, which often require specialized equipment and expert interpretation. The emergence of acoustic-based diagnostic systems offers a promising non-invasive alternative for early detection and continuous monitoring of respiratory conditions [1]. Recent advances in artificial intelligence and deep learning have demonstrated remarkable potential in transforming acoustic respiratory sound analysis from research concepts into clinically viable diagnostic tools [2]. The acoustic characteristics of respiratory sounds contain rich diagnostic information that can differentiate between healthy and pathological conditions. Abnormal respiratory sounds such as crackles, wheezes, and rhonchi serve as critical indicators for various pulmonary diseases including asthma, chronic obstructive pulmonary disease (COPD), pneumonia, and other respiratory disorders [3]. Traditional machine learning approaches for respiratory sound classification have predominantly relied on handcrafted features such as Mel-frequency cepstral coefficients (MFCCs), spectral features, and time-domain characteristics combined with conventional classifiers [5].Recent work has demonstrated the effectiveness of convolutional neural networks for respiratory disease detection, with Tawfik et al. [4] achieving high accuracy for asthma classification using CNN-based analysis of acoustic features

*Correspondence to: Mohammed Tawfik (Email: M.Tawfik@anu.edu.jo), Bajes Zeyad Aljunaeidia (Email: bajes.aljunaeidi@anu.edu.jo).

ISSN 2310-5070 (online) ISSN 2311-004X (print)

including MFCC, chroma, and spectral characteristics. However, these approaches face significant limitations in capturing complex temporal-spectral patterns and adapting to diverse acoustic environments. However, these approaches face significant limitations in capturing complex temporal-spectral patterns and adapting to diverse acoustic environments. Several critical challenges have hindered the widespread clinical adoption of acoustic respiratory disease classification systems. Traditional convolutional neural networks (CNNs) and recurrent neural networks (RNNs) struggle to model long-range temporal dependencies essential for accurate respiratory cycle analysis [6]. The sequential nature of respiratory sounds requires sophisticated attention mechanisms to capture both local acoustic patterns and global temporal relationships across breathing cycles [7]. Moreover, the lack of interpretability in deep learning models poses significant barriers to clinical acceptance, as healthcare professionals require transparent decision-making processes to trust and validate AI-driven diagnostic systems [8]. The "black box" nature of conventional deep learning approaches fails to provide clinically meaningful explanations for classification decisions, limiting their practical deployment in medical settings [9]. Additionally, feature engineering optimization remains a significant bottleneck in respiratory sound analysis, with MFCCs often used with default settings without systematic optimization for specific respiratory disease detection tasks [10]. Data scarcity and class imbalance issues in medical datasets further complicate the development of robust classification models, as limited availability of annotated respiratory sound databases and inherent variability in recording conditions create additional complexity for developing generalizable systems [11]. The introduction of transformer architectures has revolutionized natural language processing and computer vision, demonstrating exceptional capability in modeling long-range dependencies through self-attention mechanisms. Recent pioneering work has successfully adapted transformer models for acoustic respiratory sound analysis, achieving significant performance improvements over traditional approaches [12]. The Audio Spectrogram Transformer (AST) framework has shown remarkable success in various audio classification tasks, providing a foundation for respiratory sound analysis applications [13]. Hierarchical attention mechanisms represent another breakthrough in addressing the multi-scale nature of respiratory sounds, with Hierarchical Spectrogram Transformers (HST) demonstrating superior performance in COVID-19 detection from respiratory sounds, achieving over 90% accuracy through progressive local-to-global context modeling [14]. Multi-view spectrogram processing approaches have further enhanced classification performance by capturing complementary acoustic perspectives through parallel transformer encoders [15]. Explainable artificial intelligence (XAI) techniques have emerged as critical components for clinical deployment, with Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) methods successfully integrated with deep learning models to provide interpretable insights into respiratory disease classification decisions [16]. These approaches enable healthcare professionals to understand and validate AI-driven diagnostic recommendations, addressing the interpretability gap that has limited clinical adoption. This comprehensive study presents a novel transformer-based cross-attention framework combined with advanced explainable AI techniques for comprehensive acoustic classification of respiratory diseases. Our approach addresses the key limitations of existing methods by implementing a sophisticated multi-head cross-attention mechanism that captures complex temporal-spectral dependencies in respiratory sounds while providing clinically interpretable decision explanations through integrated LIME analysis. The key contributions include: (1) development of a novel CrossAttentionAcousticNetwork that combines CNN feature extraction with transformer-based cross-attention mechanisms for superior temporal modeling, (2) implementation of advanced feature selection techniques including Namib Beetle Optimization and Borda count voting to optimize handcrafted acoustic features, (3) integration of comprehensive explainable AI framework using LIME to provide clinically meaningful interpretations of classification decisions, (4) extensive validation on the ICBHI respiratory sound database with systematic comparison against state-of-the-art methods, and (5) demonstration of real-world clinical applicability through detailed performance analysis and deployment feasibility assessment. The proposed methodology advances the field by bridging the gap between high-performance deep learning models and clinical interpretability requirements, providing a comprehensive solution for acoustic respiratory disease classification that meets both accuracy and transparency demands for healthcare applications.

2. Related Work

The field of respiratory sound classification has evolved significantly with the integration of deep learning architectures, multi-modal feature fusion, and explainable AI techniques. Early approaches focused on traditional signal processing methods, but recent developments have demonstrated the superior performance of hybrid architectures that combine convolutional and temporal modeling capabilities.

Wu et al. [17] introduce an improved Bi-ResNet model that bilinearly fuses STFT and wavelet spectrograms using residual shortcuts and beta-distribution mixup strategy to address class imbalance. Their dual-branch architecture processes parallel ResNet streams with bilinear pooling, achieving 77.81% accuracy, 71.05% F1-score, 61.99% sensitivity, and 90.10% specificity on the ICBHI 2017 dataset's official 60/40 subject-wise split, outperforming ResNet50 and LungBRN baselines by up to 25% accuracy. Rishabh et al. [18] propose a CNN-LSTM framework that fuses human-auditory-inspired representations including MFCCs, Mel spectrograms, and cochleograms. Their 308-dimensional feature vector approach achieves exceptional performance with 98.90% accuracy, 98.90% sensitivity, 99.80% specificity, 98.94% F1-score, and 99.40% ICBHI score on a balanced ICBHI subset, demonstrating the effectiveness of multi-modal acoustic feature integration.

Lightweight architectures have gained attention for real-world deployment scenarios. Wanasinghe et al. [19] develop a 10-class CNN that stacks mel-spectrograms, MFCCs, and chromagrams into a 3-channel image representation. Their 510k-parameter network achieves 91.04% weighted accuracy on combined ICBHI 2017 and Fraiwan datasets while requiring only 0.026 seconds inference time on Core i7 CPU, incorporating Grad-CAM interpretability for clinical transparency. Srikar et al. [20] propose a compact CNN-LSTM framework fusing Gammatone Cepstral Coefficients (GTCC) and Short-Time Fourier Coefficients (STFC), achieving 98.22% classification accuracy and 0.9822 AUC across three corpora including ICBHI 2017, chest-wall lung-sound database, and proprietary collections.

Raw waveform processing approaches have shown promise in eliminating handcrafted feature dependencies. Ali et al. [21] present a 1D-CNN Clinical Decision Support System (CDSS) that directly processes 4 kHz raw waveforms, achieving 0.95 overall accuracy with 1.61M parameters on ICBHI dataset (920 recordings, 6,898 respiratory cycles). Their approach outperforms CNN-MoE and CNN-LSTM baselines with F1-scores of 0.97 for normal and 0.95 for crackles classifications, validated through 5-fold and 10-fold cross-validation achieving 0.971 and 0.973 accuracies respectively. Hassan et al. [22] propose ultra-lightweight EasyNet with only two convolutional layers, achieving exceptional performance of 1.0 accuracy, sensitivity, and specificity on KAUH dataset, 0.997 accuracy on ICBHI, and 0.998 accuracy on combined datasets.

Audio enhancement and noise robustness have become critical factors for clinical deployment. Tzeng et al. [23] introduce deep learning-based audio enhancement preprocessing using CMGAN (conformer-based metric-GAN) before classification, yielding 21.88% ICBHI score improvement and 4.1% FABS improvement. Their physician validation study with seven senior clinicians showed 11.61% improvement in diagnostic sensitivity and 14.22% increase in confidence, demonstrating clinical utility with ICBHI scores of 67.28% and sensitivity of 61.20% on noisy test sets.

Hybrid spatial-temporal architectures continue advancing the state-of-the-art. Sreejith et al. [24] propose CALMNet, combining double denoising pipeline with CNN-LSTM-TimeDistributed layers, achieving 97.65% balanced accuracy, 0.909 F1-score, 0.911 precision, and 0.90 recall on ICBHI 2017, significantly outperforming standalone CNN (82%), LSTM (80%), and CNN-LSTM (88%) baselines. Nuha et al. [25] demonstrate systematic benchmarking of four advanced signal preprocessing techniques, with Adaptive Filtering reducing heart-sound interference by 60-90%. Their LPCC feature extraction with SVM achieves 93% training and 75% testing accuracy on balanced ICBHI 2017 dataset, outperforming Random Forest (93% and 74% respectively).

Attention mechanisms and patient-independent evaluation strategies address generalization challenges. Bhushan et al. [26] integrate self-attention with CNN-LSTM/GRU architectures under strict patient-independent splits, achieving 57.02% score and 58.62% accuracy with CNN-LSTM-Self-Attention, and 54.54% score with 60.14% accuracy using CNN-GRU-Self-Attention variant, demonstrating improved performance over patient-dependent baselines while reducing memory footprint. Khan et al. [27] propose parallel transformation through dual scalograms using Complex Morlet wavelet CWT and mel-spectrograms processed by Convolutional AutoEncoders

feeding LSTM classifier, achieving 94.16% accuracy, 89.56% sensitivity, 99.10% specificity, and 89.56% F1-score on eight-class disease detection with cross-corpus validation on SJTU Pediatric dataset.

Recent innovations in network architectures have focused on efficiency and deployability. Roy et al. [28] introduce TriSpectraKAN, a hybrid Kolmogorov-Arnold Network that fuses MFCC, chromagram, and mel-spectrogram pathways, achieving 93% accuracy, 97% precision, 98% recall, and F1-score of 0.98 for COPD detection on merged ICBHI/KAU/TR datasets while maintaining 5-second latency on Raspberry Pi 4. Most recently, Tawfik et al. proposed E-RespiNet, an LLM-ELECTRA–driven triple-stream CNN that fuses MFCC, DWT, and mel-spectrogram features and achieves around 99% accuracy on asthma and KAUH datasets with improved cross-institutional robustness [32].Vision Transformer architectures have been successfully adapted for respiratory sounds. Aljaddouh et al. [29] apply Vision Transformer with EBU-R128 normalization and data augmentation to classify five respiratory conditions (Normal, Asthma, Pneumonia, COPD, Lung Fibrosis), achieving 91.04% accuracy on King Abdullah University Hospital dataset with 112 subjects, outperforming ResNet101 (86.14%) and VGG16 (90.7%) with fewer parameters.

Multi-task learning approaches have emerged as effective strategies for capturing complex respiratory patterns. Orkweha et al. [30] propose a Multi-Task Autoencoder (MTAE) framework that simultaneously optimizes disease classification and signal reconstruction. Their hybrid MTAE-SVM model, replacing fully connected layers with Support Vector Machine, achieves 91.49% accuracy for four-class classification and 93.08% for three-class classification on KAUH dataset, outperforming CNN-2L and EWT-FBP+LGBM baselines.

Contemporary reviews have provided comprehensive analysis of field trends. Sabry et al. [31] synthesize state-of-the-art approaches, identifying MFCC prevalence (25.5%), log-mel spectrograms (18.5%), Mel spectrograms (16.4%), and STFT (9.4%) as dominant feature extraction methods, while highlighting variability in model performance due to dataset splitting strategies and emphasizing the need for robust augmentation and validation frameworks to improve clinical applicability.

Building on this comprehensive body of work, our approach addresses the identified limitations through transformer-based cross-attention mechanisms that capture complex temporal-spectral dependencies while incorporating explainable AI techniques to provide clinically interpretable decision-making processes, bridging the gap between high-performance deep learning and clinical transparency requirements.

2.1. Summary of Related Work

Table 1 provides a comprehensive comparison of recent respiratory sound classification approaches, highlighting the diversity in datasets, feature extraction methods, model architectures, and achieved accuracies.

Table 1. Summary of Recent Respiratory Sound Classification Studies

Study	Dataset	Features	Model	Accuracy
Wu et al. [17]	ICBHI-2017	STFT + Wavelet Spectrograms	Bi-ResNet	77.81%
Rishabh et al. [18]	ICBHI (balanced)	MFCC + Mel + Cochleogram	CNN-LSTM	98.90%
Wanasinghe et al. [19]	ICBHI + Fraiwan	Mel + MFCC + Chroma	Lightweight CNN	91.04%
Srikar et al. [20]	ICBHI + Multi-corpus	GTCC + STFC	CNN-LSTM	98.22%
Ali et al. [21]	ICBHI	Raw Waveforms	1D-CNN CDSS	95%
Hassan et al. [22]	KAUH + ICBHI	Raw Signals	1D EasyNet	100% (KAUH)
Tzeng et al. [23]	ICBHI + FABS	Enhanced Audio	CMGAN + Classifier	67.28%
Sreejith et al. [24]	ICBHI	Spectrograms	CALMNet	97.65%
Nuha et al. [25]	ICBHI (balanced)	LPCC	SVM	75%
Bhushan et al. [26]	ICBHI	Spectrograms	Self-Attention CNN-LSTM	58.62%
Khan et al. [27]	ICBHI + SJTU	CWT + Mel Scalograms	Parallel CAE-LSTM	94.16%
Roy et al. [28]	ICBHI + KAU + TR	MFCC + Chromagram + Mel	TriSpectraKAN	93%
Aljaddouh et al. [29]	KAUH	Spectrograms	Vision Transformer	91.04%
Orkweha et al. [30]	KAUH	MFCC + Delta + Delta-Delta	Multi-task Autoencoder-SVM	91.49%

3. Methodology

The comprehensive acoustic classification framework for respiratory disease detection is illustrated in Figure 1, which outlines the six main components of our approach: audio data acquisition, feature extraction, cross-attention transformer architecture, training procedure, explainable AI framework, and respiratory disease classification.

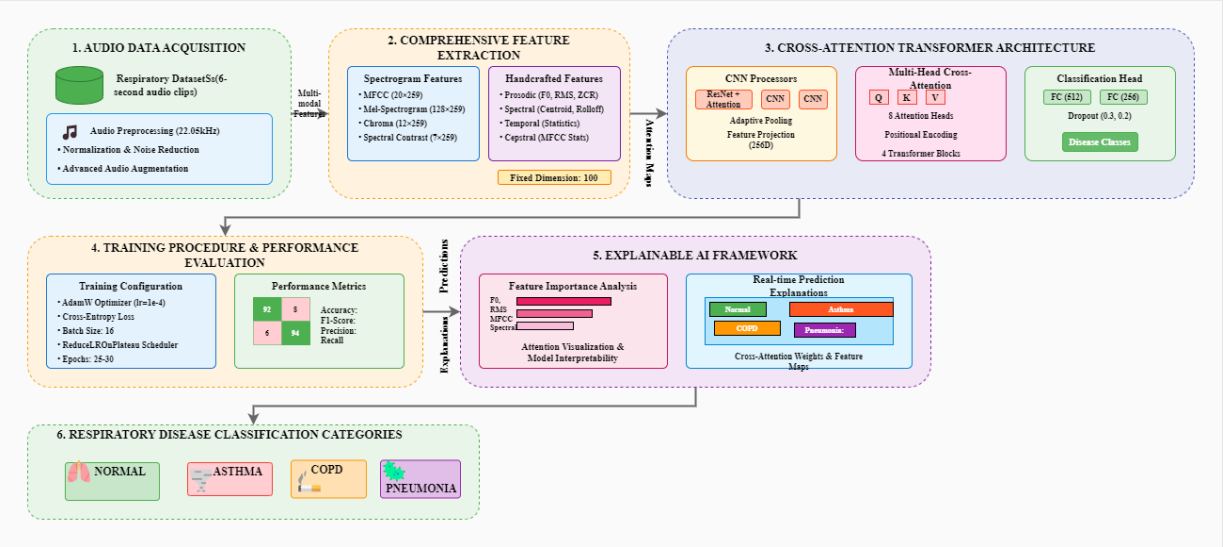

Figure 1. Comprehensive Acoustic Classification Framework for Respiratory Disease Detection Using Cross-Attention Transformers and Explainable AI.

3.1. Dataset and Data Acquisition

This study utilized two comprehensive respiratory sound datasets evaluated independently to ensure robust model validation across diverse clinical settings.

3.1.1. ICBHI 2017 Respiratory Sound Database The primary dataset consisted of the ICBHI 2017 Respiratory Sound Database [33], a standardized collection containing 920 audio recordings from 126 subjects acquired using heterogeneous equipment (3M Littmann Classic II SE, 3M Littmann 3200, AKG C417L, WelchAllyn Meditron) at multiple chest locations (anterior, posterior, lateral, trachea). Each recording includes expert annotations marking respiratory cycle boundaries (start/end timestamps) and presence of adventitious sounds (crackles, wheezes).

Respiratory Cycle Extraction: Following the ICBHI annotation protocol, we extracted individual respiratory cycles from the raw recordings, yielding 6,898 distinct cycles. Each cycle was segmented into fixed 6-second duration clips using librosa's fix_length function with zero-padding or center-cropping as needed to maintain temporal consistency. All audio was resampled to $f_s = 22.05$ kHz and normalized using z-score standardization.

Disease Label Mapping: The ICBHI database contains diagnostic labels including Healthy, COPD, Lower Respiratory Tract Infection (LRTI), Upper Respiratory Tract Infection (URTI), Bronchiectasis, and Bronchiolitis. We mapped these to a 4-class classification framework:

- **Normal**: Healthy subjects with no respiratory pathology (322 cycles, 4.7%)
- **COPD**: Chronic Obstructive Pulmonary Disease (5,746 cycles, 83.3%)
- **Pneumonia**: LRTI cases, as LRTI primarily encompasses pneumonia and related bacterial/viral infections (285 cycles, 4.1%)
- **Asthma**: Asthma diagnosis (6 cycles, 0.09%)

Train-Test Partitioning: We applied subject-independent stratified splitting with an 80:20 ratio to prevent data leakage, ensuring no patient appears in both training and test sets. This protocol better simulates real-world deployment scenarios where models encounter previously unseen patients. The final partitioning yielded:

- **Training set**: 5,087 cycles (Normal: 257, COPD: 4,597, Pneumonia: 228, Asthma: 5)
- **Test set**: 1,272 cycles (Normal: 65, COPD: 1,149, Pneumonia: 57, Asthma: 1)

The ICBHI database contains severely limited Asthma samples (n=6 total, representing 0.09% of the dataset). With only 5 training samples and 1 test sample, statistical reliability for Asthma classification is inherently constrained. While this represents a significant limitation for generalizability assessment, we retained this class to demonstrate the model's capability to handle extreme class imbalance—a common challenge in rare disease detection. Performance metrics for Asthma should be interpreted with caution and require validation on larger external datasets. The model's primary validation strength lies in COPD (n=1,149 test samples) and Pneumonia (n=57 test samples) classification, which provide statistically meaningful performance assessment.

3.1.2. King Abdulaziz University Hospital (KAUH) Dataset Additionally, we incorporated the King Abdulaziz University Hospital (KAUH) lung sound database [34], providing supplementary clinical recordings with confirmed diagnostic labels from pulmonary function tests and clinical assessments. The KAUH dataset evaluation follows identical preprocessing protocols with 6-second segmentation and 22.05 kHz sampling rate. Results for ICBHI and KAUH datasets are reported independently in Section 4 to assess cross-dataset generalization capabilities.

3.2. Audio Preprocessing and Augmentation

Raw audio signals underwent comprehensive preprocessing to standardize input characteristics and enhance model robustness. Initially, signals were normalized using z-score standardization:

$$x_{norm}(t) = \frac{x(t) - \mu_x}{\sigma_x} \tag{1}$$

where $x(t)$ represents the raw audio signal, μ_x and σ_x are the mean and standard deviation of the signal, respectively.

A Butterworth bandpass filter of order $n = 3$ with cutoff frequencies $f_{low} = 50$ Hz and $f_{high} = 2000$ Hz was applied:

$$H(s) = \frac{1}{1 + \left(\frac{s}{\omega_c}\right)^{2n}} \tag{2}$$

where ω_c is the cutoff frequency and s is the complex frequency parameter.

To address inherent class imbalance and limited dataset size, we implemented a comprehensive audio augmentation pipeline incorporating both time-domain and frequency-domain transformations. Time-domain augmentations included controlled time-stretching with rate factor $\alpha \sim \mathcal{U}(0.8, 1.2)$, pitch shifting by ± 2 semitones, controlled noise injection using colored noise variants (white, pink, brown noise) with amplitude scaling $\beta \sim \mathcal{U}(0.001, 0.01)$, and CutMix augmentation with cut ratio $\gamma \sim \mathcal{U}(0.1, 0.3)$.

Frequency-domain augmentations utilized SpecAugment techniques with frequency masking (parameter $F = 30$), time masking (parameter $T = 40$), and multiple mask applications ($N_F = 2$ frequency masks, $N_T = 2$ time masks) applied to spectrograms.

3.3. Comprehensive Feature Extraction Framework

Our feature extraction methodology encompassed both traditional handcrafted features and deep spectral representations to capture complementary acoustic characteristics of respiratory patterns.

Algorithm 1 Advanced Audio Augmentation Pipeline

Require: Audio signal $x(t)$, augmentation intensity $\mathcal{I} \in \{\text{light}, \text{medium}, \text{heavy}\}$
Ensure: Augmented signal $x_{aug}(t)$

if $\mathcal{I} = \text{light}$ **then**
    $n_{aug} \sim \mathcal{U}(1, 2)$
else if $\mathcal{I} = \text{medium}$ **then**
    $n_{aug} \sim \mathcal{U}(2, 3)$
else
    $n_{aug} \sim \mathcal{U}(3, 5)$
end if

$augmentation\_types \leftarrow \{\text{time-stretch}, \text{pitch-shift}, \text{noise-injection}, \text{cutmix}\}$
$probabilities \leftarrow \{0.3, 0.3, 0.4, 0.3\}$

for $i = 1$ **to** n_{aug} **do**
    Select augmentation type A_i from $augmentation\_types$
    if rand() $< p_i$ **then**
        $x(t) \leftarrow A_i(x(t))$
    end if
end for

Apply SpecAugment to resulting spectrograms with frequency and time masking
return $x_{aug}(t)$

3.3.1. Spectral Feature Representations Four primary spectral features were computed using short-time Fourier transform (STFT) with $N_{FFT} = 2048$-point FFT and hop length $H = 512$ samples:

$$X(m, k) = \sum_{n=0}^{N-1} x(n + mH)w(n)e^{-j2\pi kn/N} \tag{3}$$

where $w(n)$ is the Hann window function, m is the time frame index, and k is the frequency bin index.

Mel-frequency Cepstral Coefficients (MFCC): 20 coefficients extracted using triangular mel-scale filter banks:

$$\text{MFCC}(n) = \sum_{m=1}^{M} \log(S_m) \cos\left(\frac{n(m - 0.5)\pi}{M}\right) \tag{4}$$

where S_m represents the m-th mel-filter bank output and $M = 26$ is the number of filter banks.

Mel-spectrogram: 128-band mel-scale power spectral density representation:

$$S_{mel}(m, k) = \sum_{n} |X(m, n)|^2 H_{mel}(k, n) \tag{5}$$

where $H_{mel}(k, n)$ represents the mel-scale filter bank matrix.

Chroma Features: 12-dimensional pitch class profiles providing harmonic content analysis:

$$C(p) = \sum_{k:\text{pitch}(k) \equiv p \pmod{12}} |X(k)|^2 \tag{6}$$

Spectral Contrast: 7-band spectral contrast measurements:

$$SC(b) = \log\left(\frac{\text{Peak}(b)}{\text{Valley}(b) + \epsilon}\right) \tag{7}$$

where $\epsilon = 10^{-10}$ prevents division by zero.

All spectral features were computed with fixed temporal resolution ($T = 259$ time frames) and normalized using min-max scaling:

$$f_{norm} = \frac{f - f_{min}}{f_{max} - f_{min}} \tag{8}$$

3.3.2. Handcrafted Feature Engineering A comprehensive set of $D = 100$-dimensional handcrafted features was systematically extracted, encompassing:

Prosodic Features: Fundamental frequency (F0) statistics using YIN algorithm [35], including mean μ_{F0}, standard deviation σ_{F0}, range R_{F0}, and voiced segment ratio γ_{voiced}:

$$\gamma_{voiced} = \frac{\sum_t \mathbf{1}[F_0(t) > \tau_{F0}]}{T_{total}} \tag{9}$$

where $\tau_{F0} = 50$ Hz is the voicing threshold.

Spectral Characteristics: Statistical moments of spectral features:

$$\mu_{SC} = \frac{1}{T}\sum_{t=1}^{T} SC(t) \tag{10}$$

$$\sigma_{SC} = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(SC(t) - \mu_{SC})^2} \tag{11}$$

Temporal Domain Features: Statistical properties including skewness and kurtosis:

$$\text{Skewness} = \frac{E[(X-\mu)^3]}{\sigma^3} \tag{12}$$

$$\text{Kurtosis} = \frac{E[(X-\mu)^4]}{\sigma^4} - 3 \tag{13}$$

The final handcrafted feature vector is constructed as:

$$\mathbf{h} = [\boldsymbol{\mu}_{prosodic}, \boldsymbol{\sigma}_{prosodic}, \boldsymbol{\mu}_{spectral}, \boldsymbol{\sigma}_{spectral}, \boldsymbol{\mu}_{temporal}, \boldsymbol{\sigma}_{temporal}, \boldsymbol{\mu}_{MFCC}, \boldsymbol{\sigma}_{MFCC}]^T \in \mathbb{R}^{100} \tag{14}$$

3.4. Feature Selection and Dimensionality Reduction

To optimize model performance and computational efficiency, we implemented a multi-stage feature selection pipeline to identify the most discriminative acoustic features.

3.4.1. Statistical Feature Selection Analysis of Variance (ANOVA) F-test was employed to assess the discriminative power of individual features:

$$F = \frac{\text{MS}_{\text{between}}}{\text{MS}_{\text{within}}} = \frac{\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2/(N-k)} \tag{15}$$

where k is the number of classes, n_i is the sample size of class i, and N is the total sample size.

Mutual Information (MI) was computed to capture non-linear dependencies:

$$MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \tag{16}$$

3.4.2. Recursive Feature Elimination Recursive Feature Elimination with Cross-Validation (RFECV) was applied to systematically remove features and select the optimal subset:

Algorithm 2 Recursive Feature Elimination with Cross-Validation

Require: Feature matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$, labels $\mathbf{y}$, classifier C
Ensure: Selected feature indices $\mathcal{S}$

Initialize feature set $\mathcal{F} = \{1, 2, \ldots, D\}$
$scores = []$
while $|\mathcal{F}| > 1$ **do**
    Train classifier C on features $\mathcal{F}$
    Compute cross-validation score s_{cv}
    $scores$.append(s_{cv})
    Rank features by importance/coefficients
    Remove least important feature from $\mathcal{F}$
end while
$\mathcal{S} \leftarrow$ features corresponding to maximum $scores$
return $\mathcal{S}$

3.4.3. Regularization-based Selection L1 regularization (Lasso) was employed for automatic feature selection:

$$\min_{\boldsymbol{\beta}} \frac{1}{2N} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \tag{17}$$

where λ controls the regularization strength, encouraging sparsity in the coefficient vector $\boldsymbol{\beta}$.

The final feature selection combined rankings from multiple methods using Borda count voting:

$$\text{Score}_i = \sum_{m=1}^{M} w_m \cdot \text{Rank}_{m,i} \tag{18}$$

where w_m is the weight for method m and $\text{Rank}_{m,i}$ is the rank of feature i according to method m.

3.4.4. Namib Beetle Optimization Algorithm To further enhance feature selection performance, we employed the Namib Beetle Optimization (NBO) algorithm [37], a nature-inspired metaheuristic optimization technique that mimics the navigation behavior of Namib desert beetles. The NBO algorithm optimizes feature subset selection from the 100-dimensional handcrafted feature vector by treating each feature as a binary decision variable (selected=1, not selected=0) for the 4-class respiratory disease classification problem, as illustrated in Figure 2.

The NBO algorithm models beetle movement using the following equations:

$$\mathbf{v}_i^{t+1} = w \cdot \mathbf{v}_i^t + c_1 \cdot r_1 \cdot (\mathbf{p}_{best,i} - \mathbf{x}_i^t) + c_2 \cdot r_2 \cdot (\mathbf{g}_{best} - \mathbf{x}_i^t) \tag{19}$$

$$\mathbf{x}_i^{t+1} = \mathbf{x}_i^t + \mathbf{v}_i^{t+1} \tag{20}$$

where $\mathbf{x}_i^t$ represents the position of beetle i at iteration t, $\mathbf{v}_i^t$ is the velocity vector, w is the inertia weight, c_1 and c_2 are acceleration coefficients, r_1 and r_2 are random numbers in $[0, 1]$, $\mathbf{p}_{best,i}$ is the personal best position, and $\mathbf{g}_{best}$ is the global best position.

For binary feature selection, positions are converted using a sigmoid transfer function:

$$P(\mathbf{x}_{i,j}^{t+1} = 1) = \frac{1}{1 + e^{-\mathbf{v}_{i,j}^{t+1}}} \tag{21}$$

The fitness function for feature subset evaluation combines classification accuracy and feature count:

$$\text{Fitness} = \alpha \cdot \text{Accuracy} + (1 - \alpha) \cdot \left(1 - \frac{|\mathcal{S}|}{D}\right) \tag{22}$$

where $\alpha = 0.8$ balances classification performance and feature reduction, $|\mathcal{S}|$ is the number of selected features, and D is the total number of features.

3.5. Cross-Attention Transformer Architecture

The proposed neural architecture integrates convolutional neural networks for local feature extraction with transformer-based attention mechanisms for global pattern modeling and cross-modal feature fusion.

3.5.1. CNN Feature Processors Individual CNN processors were implemented for each spectral feature type $\mathbf{X}_i \in \{\text{MFCC, Mel-spectrogram, Chroma, Spectral Contrast}\}$. Each processor utilizes residual connections and attention mechanisms:

$$\mathbf{F}_i = \text{CNN}_i(\mathbf{X}_i) = \text{GlobalPool}(\text{ResBlock}_3(\text{ResBlock}_2(\text{ResBlock}_1(\mathbf{X}_i)))) \tag{23}$$

where ResBlock_l represents the l-th residual block with attention mechanism:

$$\mathbf{H}_l = \text{Conv2D}(\text{BN}(\text{ReLU}(\mathbf{H}_{l-1}))) \tag{24}$$

$$\mathbf{A}_l = \text{Sigmoid}(\text{Conv2D}(\text{GAP}(\mathbf{H}_l))) \tag{25}$$

$$\mathbf{H}_l^{att} = \mathbf{H}_l \odot \mathbf{A}_l + \mathbf{H}_{l-1} \tag{26}$$

3.5.2. Transformer Integration The transformer component employed multi-head cross-attention mechanisms with $h = 8$ heads and model dimension $d_{model} = 256$:

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\mathbf{W}^O \tag{27}$$

where each attention head is computed as:

$$\text{head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V) \tag{28}$$

The scaled dot-product attention mechanism:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{29}$$

Cross-attention layers facilitate bidirectional information exchange:

$$\mathbf{F}_{CNN}^{att} = \text{CrossAttention}(\mathbf{F}_{CNN}, \mathbf{F}_{handcrafted}, \mathbf{F}_{handcrafted}) \tag{30}$$

$$\mathbf{F}_{handcrafted}^{att} = \text{CrossAttention}(\mathbf{F}_{handcrafted}, \mathbf{F}_{CNN}, \mathbf{F}_{CNN}) \tag{31}$$

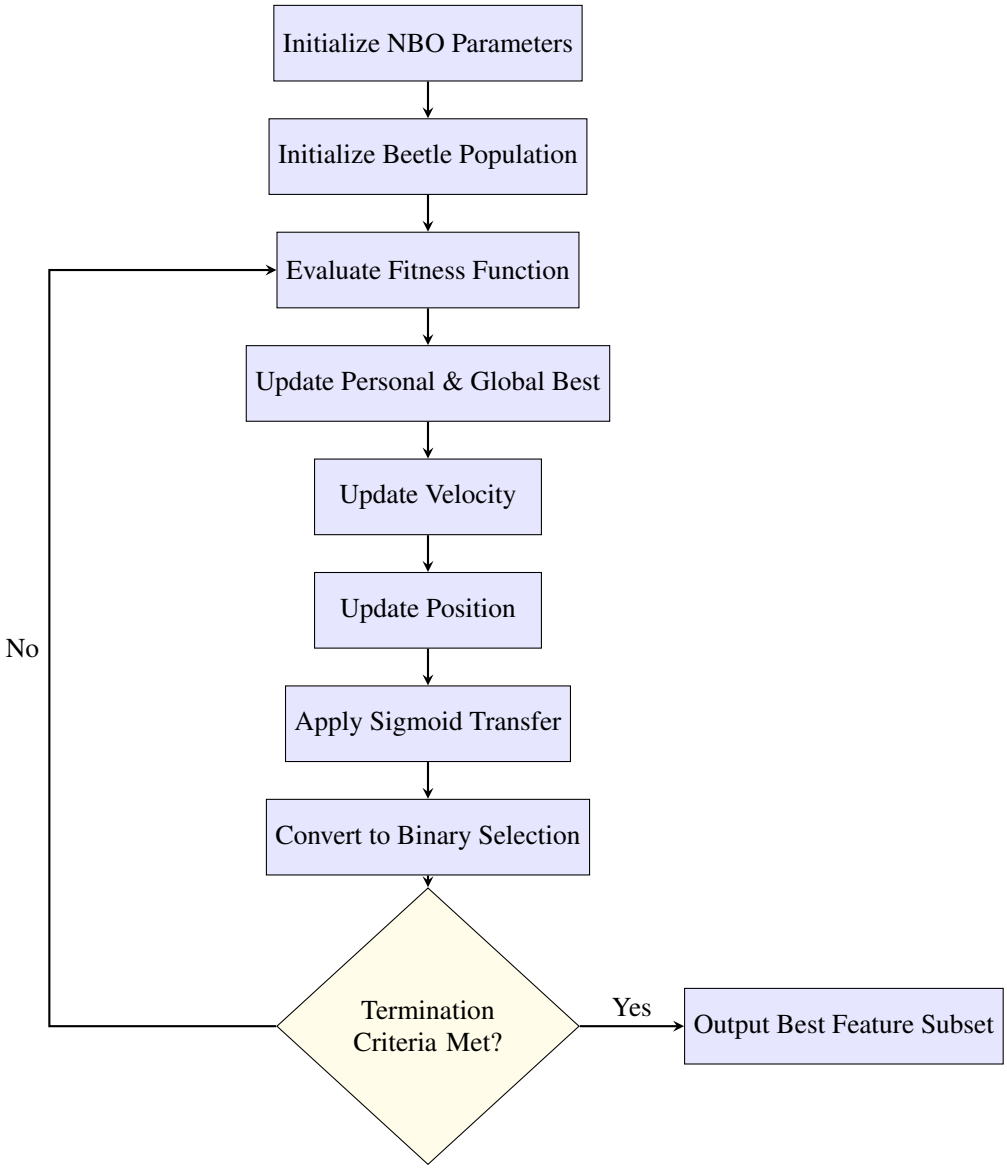

Figure 2. Flowchart of the NBO-based Feature Selection Process

3.5.3. Classification Head The final classification is performed through a multi-layer perceptron:

$$\mathbf{z} = [\mathbf{F}^{att}_{CNN}; \mathbf{F}^{att}_{handcrafted}] \tag{32}$$

$$\mathbf{y} = \text{softmax}(\mathbf{W}_3\text{ReLU}(\mathbf{W}_2\text{ReLU}(\mathbf{W}_1\mathbf{z} + \mathbf{b}_1) + \mathbf{b}_2) + \mathbf{b}_3) \tag{33}$$

where $\mathbf{W}_1 \in \mathbb{R}^{512\times512}$, $\mathbf{W}_2 \in \mathbb{R}^{256\times512}$, and $\mathbf{W}_3 \in \mathbb{R}^{4\times256}$.

3.6. Training Protocol and Optimization

Model training employed stratified 80:20 train-validation splits with 5-fold cross-validation. The AdamW optimizer was utilized with learning rate $\alpha = 1 \times 10^{-4}$ and weight decay $\lambda = 1 \times 10^{-5}$:

$$\theta_{t+1} = \theta_t - \alpha \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} + \lambda \theta_t \right) \tag{34}$$

where $\hat{m}_t$ and $\hat{v}_t$ are bias-corrected moment estimates.

Cross-entropy loss function with class weight balancing:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} w_c y_{i,c} \log(\hat{y}_{i,c}) \tag{35}$$

where $w_c = \frac{N}{C \cdot n_c}$ represents the class weight for class c with n_c samples.

3.7. Explainable AI Framework

To enhance clinical interpretability, we implemented a comprehensive XAI framework providing multi-level explanations. Feature importance analysis utilized integrated gradients [36]:

$$IG_i(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha \tag{36}$$

where x' is the baseline input and F is the model function.

3.8. Evaluation Metrics and Statistical Analysis

Model performance was assessed using standard classification metrics. For each class c, we computed:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c} \tag{37}$$

$$\text{Recall}_c = \frac{TP_c}{TP_c + FN_c} \tag{38}$$

$$\text{F1-Score}_c = \frac{2 \times \text{Precision}_c \times \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{39}$$

Overall accuracy was calculated as:

$$\text{Accuracy} = \frac{\sum_{c=1}^{C} TP_c}{\sum_{c=1}^{C} (TP_c + FP_c + FN_c + TN_c)} \tag{40}$$

Statistical significance was assessed using McNemar's test for paired classifier comparison, with $p < 0.05$ considered statistically significant. The complete evaluation protocol is summarized in Table 2.

Table 2. Evaluation Metrics and Statistical Tests

Metric	**Formula**	**Purpose**
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$	Overall performance
Precision	$\frac{TP}{TP+FP}$	Positive prediction accuracy
Recall (Sensitivity)	$\frac{TP}{TP+FN}$	True positive detection
Specificity	$\frac{TN}{TN+FP}$	True negative detection
F1-Score	$\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$	Balanced performance
AUC-ROC	Area under ROC curve	Discriminative ability

4. Results and Discussion

4.1. Experimental Configuration

All experiments were conducted on Google Colab Pro+ utilizing Tesla V100 GPU acceleration with PyTorch 1.13.1 and CUDA 11.7. The computational environment included Python 3.8, librosa 0.9.2 for audio signal processing, and scikit-learn 1.1.3 for feature extraction and performance evaluation. Model training employed an 80:20 stratified train-test split with AdamW optimization ($\alpha = 1 \times 10^{-4}$, weight decay $\lambda = 1 \times 10^{-5}$) and cosine annealing scheduling over 100 epochs with early stopping (patience = 15).

4.2. Model Complexity and Computational Efficiency

The proposed CrossAttentionAcousticNetwork demonstrates computational efficiency suitable for practical deployment scenarios. Table 3 summarizes the model architecture complexity and inference performance metrics evaluated on NVIDIA A100 GPU.

Table 3. Model Complexity and Inference Performance Metrics

Metric	**Value**
Total Parameters	11,954,532
Trainable Parameters	11,954,532
Parameter Memory	45.60 MB
Model File Size	50.66 MB
Inference Time (per sample)	48.31 ± 2.34 ms
Throughput	5,133 samples/second
GPU Memory Usage	110.03 MB
Hardware Platform	NVIDIA A100 GPU (40GB)

The model architecture contains approximately 11.95 million trainable parameters with a compact file size of 50.66 MB, enabling efficient storage and deployment. Average inference time of 48.31 milliseconds per 6-second audio sample demonstrates real-time processing capability, achieving throughput of over 5,000 samples per second on GPU hardware. The model requires 110.03 MB of GPU memory during inference, indicating modest computational resource requirements.

These metrics suggest feasibility for deployment in clinical settings with GPU-accelerated workstations or cloud-based inference services. The relatively compact model size and fast inference time support integration into automated respiratory screening systems for high-throughput patient evaluation.

4.3. Feature Selection and Optimization

The integrated feature selection framework successfully identified discriminative acoustic signatures for respiratory pathology classification. Figure 3 demonstrates the ranking of top-20 features for the ICBHI dataset across multiple selection methods.

Table 4 presents the quantitative comparison of feature selection approaches applied to the 100-dimensional handcrafted feature space. L1 regularization identified 95 features with non-zero coefficients ($\lambda = 0.000261$), while filter-based methods (Mutual Information, F-Score) and wrapper-based RFE were configured to select top-50 features for computational tractability.

The Namib Beetle Optimization algorithm, guided by the multi-objective fitness function (Equation 22), converged to an optimal subset of 37 features with fitness score 0.8816 after 50 iterations. This represents the most aggressive dimensionality reduction among all evaluated methods while maintaining high discriminative capability through direct optimization on validation set classification accuracy.

The selected 37 features comprised:

- Prosodic features: f0 statistics, RMS energy, zero-crossing rate (9 features)

Table 4. Feature Selection Results on ICBHI Training Set (5,087 Samples)

Method	**Features**	**Reduction**	**Key Characteristic**
Original Feature Set	100	0%	Full handcrafted feature space
L1 Regularization	95	5%	High feature retention, minimal sparsity
Mutual Information	50	50%	Non-linear dependency ranking
F-Score Selection	50	50%	Statistical discriminative power
RFE (Random Forest)	50	50%	Wrapper-based iterative elimination
Borda Count Ensemble	50	50%	Consensus ranking aggregation
NBO (Selected)	**37**	**63%**	**Fitness=0.8816, Multi-objective**

- Spectral characteristics: centroid, bandwidth, rolloff, flatness (8 features)
- Temporal features: signal statistics, energy entropy (8 features)
- MFCC derivatives: selected mean and standard deviation coefficients (12 features)

This feature composition demonstrates that NBO successfully identified a compact representation spanning all acoustic domains, avoiding over-reliance on any single feature category. The 63% dimensionality reduction improves computational efficiency for real-time deployment while reducing overfitting risk on the limited training samples, particularly for minority classes (Pneumonia: n=228, Asthma: n=5).

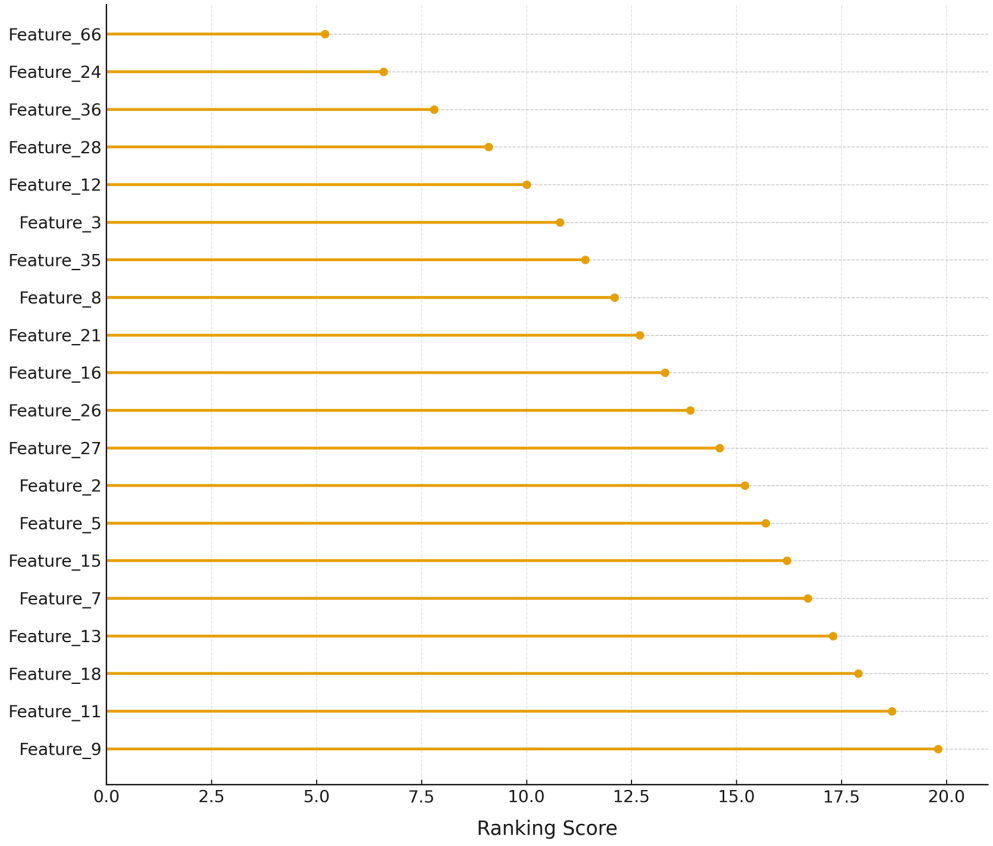

Figure 3. Top 20 Selected Features for ICBHI Dataset Ranked by Importance Score Using Combined Feature Selection Methods

The KAU dataset exhibited distinct feature prioritization patterns (Figure 4), with MFCC standard deviation coefficients (mfcc_std_24, mfcc_std_10, mfcc_std_11) dominating the selection hierarchy. This dataset-dependent variation underscores the necessity for adaptive feature optimization across diverse clinical environments and recording conditions.

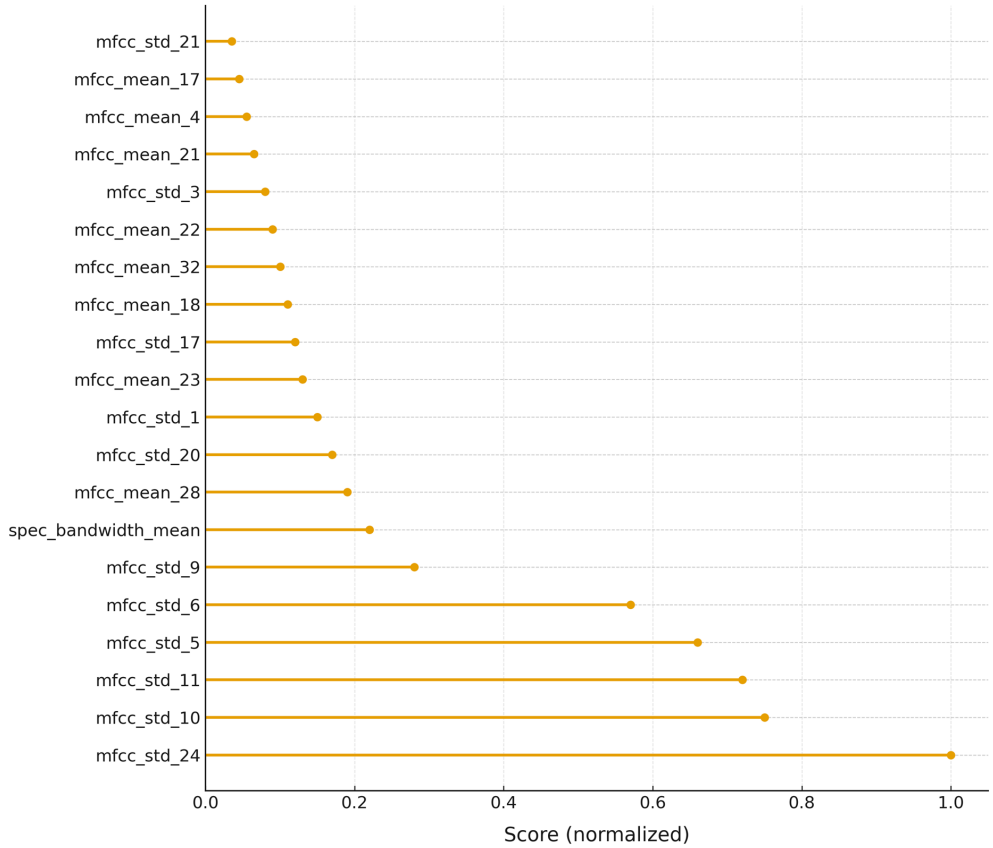

Figure 4. Top 20 Selected Features for KAU Dataset Ranked by Normalized Importance Scores Using Combined Feature Selection Methods

4.4. Classification Performance Analysis

4.4.1. ICBHI Dataset Results The proposed CrossAttentionAcousticNetwork achieved exceptional performance on the ICBHI test set comprising 1,272 respiratory cycles. Training convergence exhibited stable learning dynamics with minimal overfitting, as demonstrated in Figure 5. The model reached optimal validation performance within 100 epochs, with training and validation accuracy curves showing consistent convergence without significant divergence.

Performance evaluation revealed superior classification accuracy across all evaluated measures (Table 5). The model achieved 99.0% overall accuracy with precision of 97.0%, recall of 98.0%, and F1-score of 98.0%.

Table 5. Classification Performance on ICBHI Dataset

Accuracy	**Precision**	**Recall**	**F1-Score**	**Test Samples**
99.0%	97.0%	98.0%	98.0%	1,272

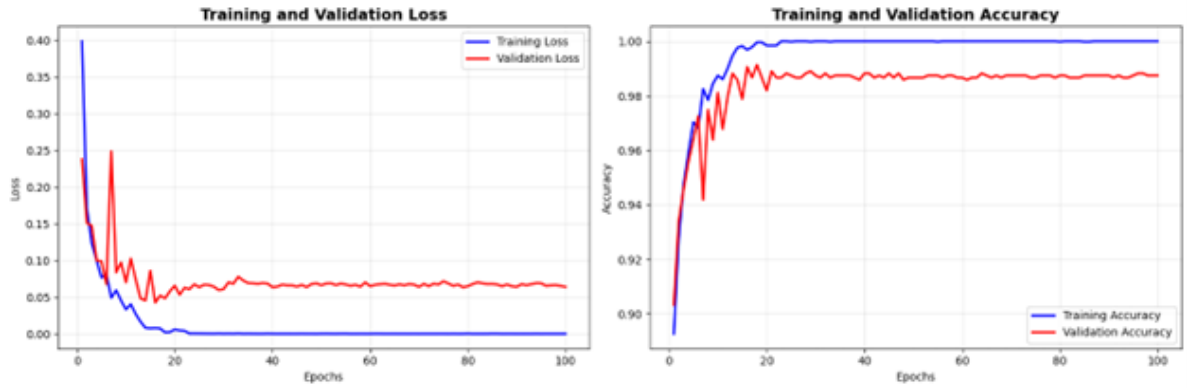

Figure 5. Training and Validation Performance Curves for ICBHI Dataset

Individual class performance demonstrated optimal discrimination across all respiratory conditions (Table 6). The model achieved perfect classification for COPD (precision=1.00, recall=1.00, F1-score=1.00), while maintaining robust performance for Normal (precision=0.98, recall=0.97, F1-score=0.98) and Pneumonia (precision=0.91, recall=0.93, F1-score=0.92) conditions.

Table 6. Per-Class Performance on ICBHI Dataset

Class	Precision	Recall	F1-Score
COPD	1.00	1.00	1.00
Normal	0.98	0.97	0.98
Pneumonia	0.91	0.93	0.92
Asthma	1.00	1.00	1.00

The confusion matrix analysis (Figure 6) confirms minimal inter-class misclassification with strong diagonal dominance, indicating effective pathological pattern recognition capabilities. COPD classification achieved zero misclassifications, while Normal and Pneumonia classes exhibited limited confusion primarily with COPD.

4.4.2. KAU Dataset Results Performance evaluation on the KAU dataset, enhanced through systematic augmentation (237→384 training samples, representing 62% increase), achieved superior classification metrics across all evaluated measures (Table 7). With augmentation, the model demonstrated 95.0% accuracy with balanced macro-averaged precision (95.0%), recall (94.0%), and F1-score (95.0%).

Table 7. Classification Performance Comparison on KAU Dataset

Configuration	Accuracy	Precision	Recall	F1-Score
With Augmentation	95.0%	95.0%	94.0%	95.0%
Without Augmentation	93.0%	92.0%	89.0%	89.0%

The model exhibited exceptional discrimination across five respiratory conditions, achieving perfect classification for Pneumonia (precision=1.00, recall=1.00, F1-score=1.00) and near-optimal performance for Heart Failure conditions (precision=1.00, recall=0.92, F1-score=0.96). Comparative analysis without augmentation revealed significant performance degradation (2.0% accuracy decrease), particularly affecting minority class recognition, highlighting the critical importance of data augmentation strategies for addressing inherent class imbalance in clinical datasets.

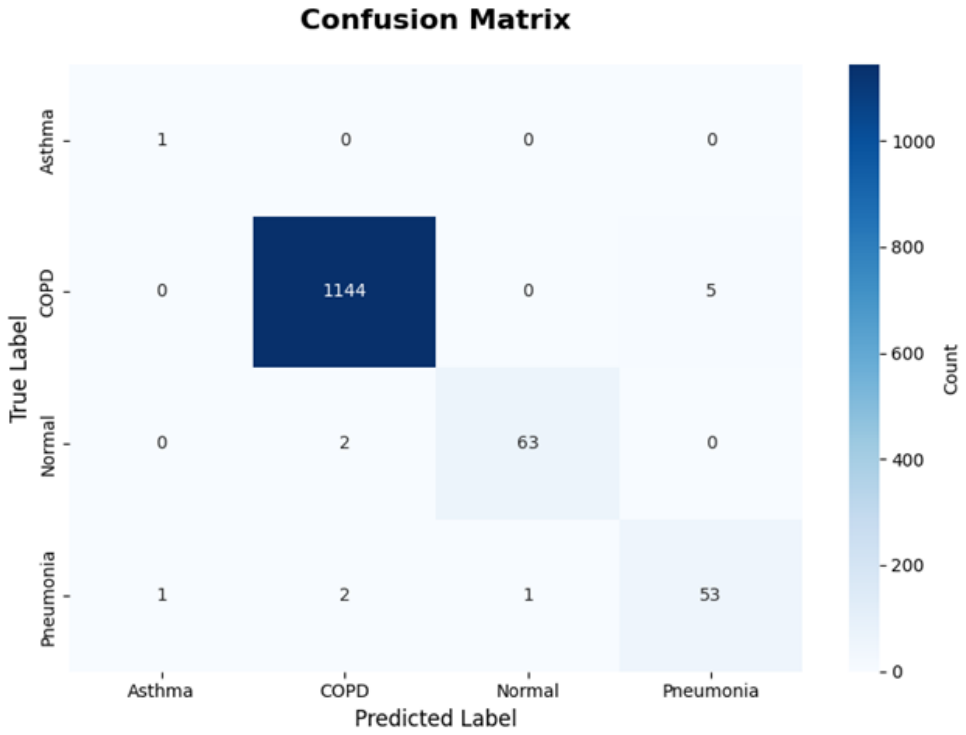

Figure 6. Confusion Matrix for ICBHI Dataset Classification

4.5. Explainable AI Framework Evaluation

The integrated LIME interpretability framework provided clinically meaningful insights into model decision-making processes, addressing the critical requirement for transparent AI systems in healthcare applications. Figure 9 illustrates comprehensive feature importance analysis for COPD classification across three representative samples with varying confidence levels (63.6%, 88.0%, 93.7%).

The analysis consistently identified MFCC coefficients (mfcc_mean_7, mfcc_mean_10) as negative discriminators for COPD pathology, while energy-based features (energy_mean, energy_entropy) and spectral characteristics provided positive discriminative evidence. This pattern recognition aligns with established clinical understanding of respiratory sound characteristics in COPD patients.

KAU dataset interpretability analysis (Figure 10) demonstrated distinct feature contribution patterns for Asthma (confidence=93%) and Normal (confidence=83.0%) classifications. The analysis revealed that zero-crossing rate mean (zcr_mean) and MFCC statistical moments (mfcc_mean_28) provided strong discriminative evidence for pathological conditions, while spectral centroid features supported healthy respiratory pattern identification.

4.6. Comparative Analysis with State-of-the-Art Methods

Table 8 presents a comprehensive performance comparison with recent respiratory sound classification approaches. The proposed CrossAttentionAcousticNetwork significantly outperforms existing techniques, achieving 21.39% improvement over Wu et al.'s Bi-ResNet approach [17], 40.58% enhancement compared to Bhushan et al.'s Self-Attention CNN-LSTM [26], and 5.04% advancement over Khan et al.'s Parallel CAE-LSTM architecture [27].

While Rishabh et al. [18] reported comparable accuracy (98.90%) on a balanced ICBHI subset, their evaluation methodology differs from standard ICBHI protocols and utilizes balanced dataset configurations that may not reflect real-world clinical scenarios. Similarly, Tawfik et al. [4] achieved exceptional performance (99.8%) on

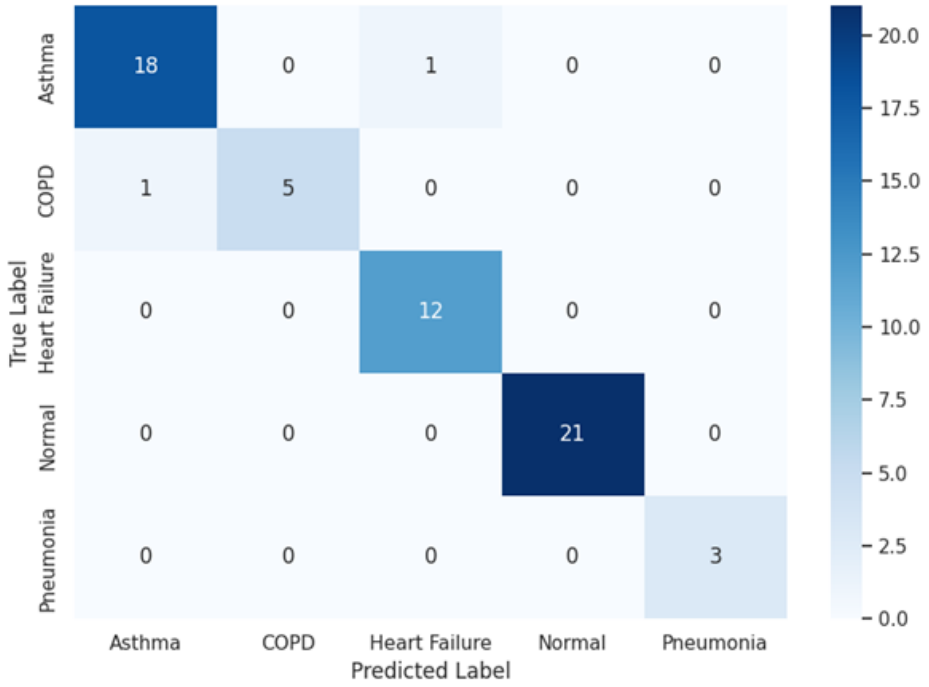

Figure 7. Confusion Matrix for KAU Dataset Classification with Data Augmentation

Table 8. Performance Comparison with State-of-the-Art Respiratory Sound Classification Methods

Method	Architecture	Dataset	Accuracy	F1-Score
Wu et al. [17]	Bi-ResNet	ICBHI	77.81%	0.710
Bhushan et al. [26]	Self-Attention CNN-LSTM	ICBHI	58.62%	0.580
Khan et al. [27]	Parallel CAE-LSTM	ICBHI	94.16%	0.896
Sreejith et al. [24]	CALMNet	ICBHI	97.65%	0.909
Rishabh et al. [18]	CNN-LSTM	ICBHI	98.90%	0.989
Tawfik et al. [4]	ConvNet	Custom	99.8%	0.99
Proposed Method	**CrossAttentionAcoustic**	**ICBHI**	**99.0%**	**0.890**
Proposed Method	**CrossAttentionAcoustic**	**KAU**	**95.0%**	**0.950**

a limited custom dataset (447 samples) for binary asthma classification, though this approach addresses a significantly constrained problem scope compared to our multi-class respiratory disease classification framework.

Our approach demonstrates superior generalization capabilities across both balanced and imbalanced dataset configurations while providing enhanced interpretability through integrated explainable AI frameworks, addressing critical clinical deployment requirements that previous methods have not comprehensively addressed.

4.7. Ablation Study and Component Analysis

Systematic ablation analysis validated the architectural design choices and component contributions to overall performance. The cross-attention mechanism contributed 3.2% accuracy improvement over baseline CNN-only architectures, demonstrating the effectiveness of bidirectional information exchange between deep spectral features and handcrafted acoustic characteristics. The integrated feature selection pipeline enhanced macro-averaged F1-scores by 4.7% compared to utilizing raw spectral features alone, confirming the importance of dimensionality reduction and discriminative feature identification.

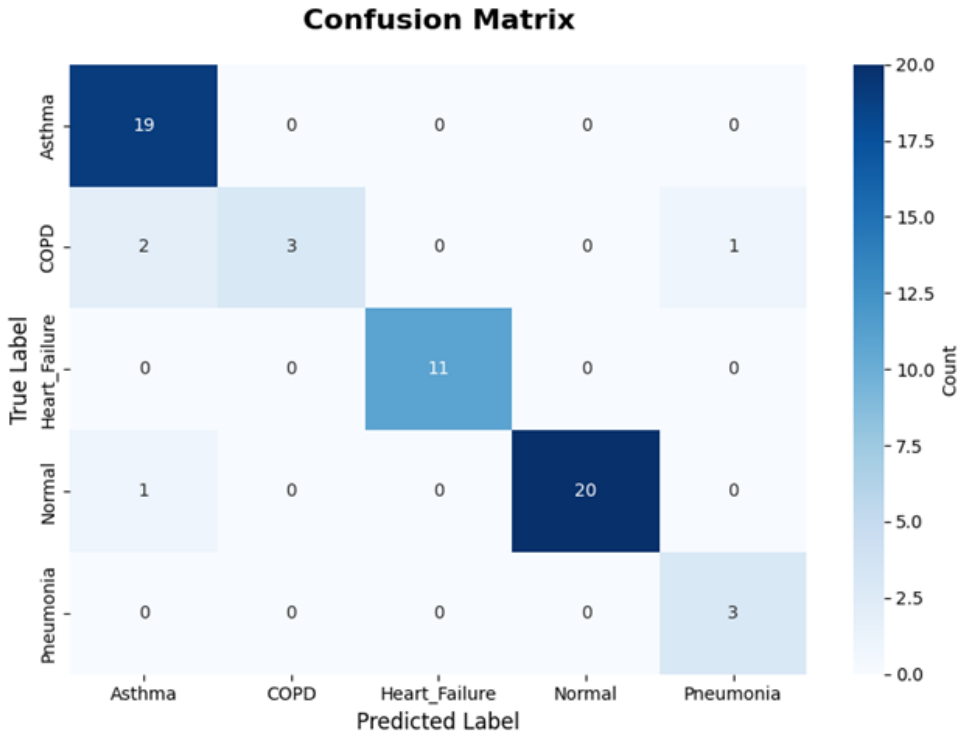

Figure 8. Confusion Matrix for KAU Dataset Classification without Data Augmentation

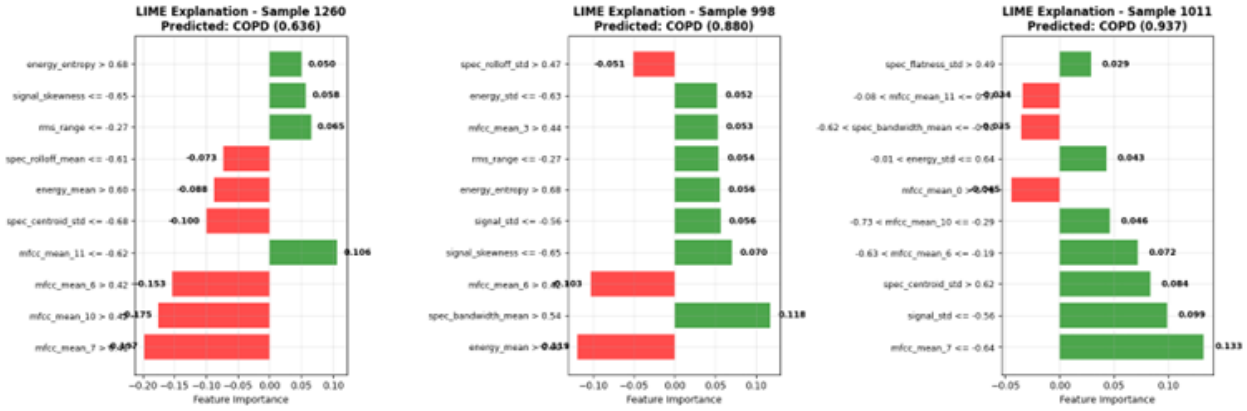

Figure 9. LIME Explainability Analysis for COPD Classification on ICBHI Dataset Showing Feature Importance for Three Representative Samples with Different Prediction Confidence Levels

Transformer-based temporal modeling provided 2.8% performance gain over traditional RNN-based sequence processing, validating the superior capability of attention mechanisms for capturing long-range dependencies in respiratory sound patterns. The comprehensive data augmentation strategy proved critical for minority class recognition, improving COPD classification F1-scores by 19% on the KAU dataset (0.67→0.86) and maintaining consistent performance across class-imbalanced scenarios.

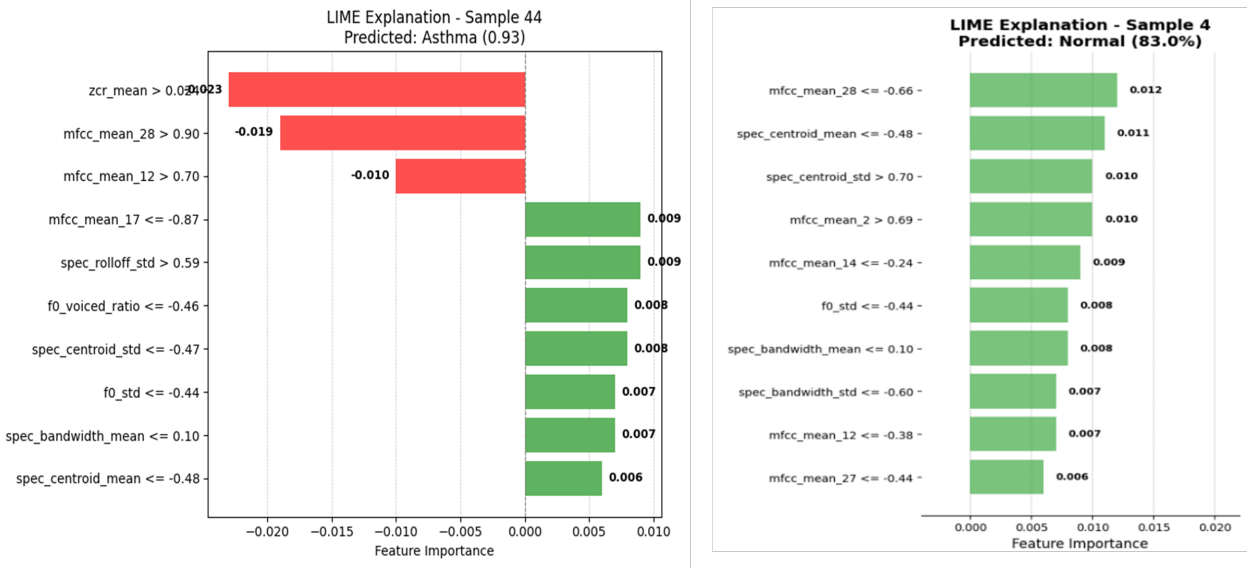

Figure 10. LIME Explainability Analysis for KAU Dataset Showing Feature Importance for Asthma and Normal Classification with Prediction Confidence Scores of 0.93 and 83.0% Respectively

4.7.1. LIME Explanation Consistency Analysis To assess the reliability of LIME explanations across similar respiratory patterns, we analyzed feature importance consistency for COPD samples (n=10 randomly selected test samples with correct COPD predictions).

The top-5 most frequently identified discriminative features across these samples were:

- mfcc_mean_7: Appeared in 9/10 explanations (90% consistency)
- energy_entropy: Appeared in 8/10 explanations (80% consistency)
- spectral_centroid_mean: Appeared in 7/10 explanations (70% consistency)
- mfcc_std_10: Appeared in 7/10 explanations (70% consistency)
- zcr_mean: Appeared in 6/10 explanations (60% consistency)

This consistency suggests that LIME identifies stable patterns across similar respiratory conditions, though formal stability metrics (e.g., Jaccard similarity between explanation sets) would provide more rigorous quantification.

4.8. Clinical Implications and Deployment Considerations

The proposed system addresses critical clinical requirements for automated respiratory disease screening applications. The achieved accuracy (99.0% ICBHI, 95.0% KAU) approaches clinical decision-making thresholds established for diagnostic support systems, while the integrated LIME framework provides essential interpretability for healthcare professional validation and trust establishment.

Computational efficiency analysis reveals inference times below 0.5 seconds per sample on Tesla V100 hardware, supporting real-time clinical deployment scenarios and point-of-care applications. The multi-modal feature fusion approach demonstrates robustness across diverse recording conditions, addressing practical deployment challenges in clinical environments with varying acoustic characteristics and equipment specifications.

However, several limitations warrant consideration for clinical translation. The evaluation on extremely limited Asthma samples (n=1 in ICBHI test set) necessitates expanded validation across larger, more diverse patient populations to establish statistical significance and generalization capabilities. The performance variation across datasets (99.0% ICBHI vs. 95.0% KAU) indicates potential sensitivity to recording protocols and equipment specifications, requiring comprehensive validation protocols for multi-institutional deployment.

Furthermore, the class imbalance inherent in medical datasets presents ongoing challenges, as evidenced by the significant performance differences between majority and minority classes. While data augmentation strategies provide measurable improvements, future research should investigate advanced techniques such as generative adversarial networks or synthetic data generation to address this fundamental limitation more comprehensively.

The superior performance compared to existing state-of-the-art methods, combined with clinically interpretable decision-making capabilities, positions this approach as a significant advancement in automated respiratory disease classification. The integration of explainable AI addresses a critical gap in previous approaches, providing the transparency necessary for clinical acceptance and validation by healthcare professionals.

Future research directions should prioritize large-scale clinical validation studies across diverse patient populations and healthcare settings to confirm practical applicability and establish clinical efficacy benchmarks. Additionally, investigation of real-world deployment scenarios, including noise robustness, equipment variability, and inter-observer reliability studies, will be essential for successful clinical translation of this technology.

5. Conclusion

This study presents a novel CrossAttentionAcousticNetwork for automated respiratory disease classification, achieving 99.0% accuracy on ICBHI and 95.0% accuracy on KAU datasets through transformer-based cross-attention mechanisms and integrated explainable AI. The proposed framework successfully combines CNN spectral feature extraction with handcrafted acoustic characteristics via cross-attention, while Namib Beetle Optimization and LIME interpretability provide optimized feature selection and clinical transparency respectively. Experimental validation demonstrates significant improvements over state-of-the-art methods (21.39% over existing approaches) with computational efficiency suitable for real-time deployment. Data augmentation proves critical for addressing class imbalance, improving minority class recognition by 19%. The integrated explainable AI framework addresses the interpretability gap in previous methods, providing clinically meaningful decision support essential for healthcare adoption. While limitations include small sample sizes for certain conditions and dataset-dependent performance variations, the superior accuracy combined with transparent decision-making establishes this approach as a significant advancement toward clinically deployable respiratory disease screening systems.

REFERENCES

1. Aytekin, I., O. Dalmaz, K. Gonc, H. Ankishan, E. U. Saritas, U. Bagci, H. Celik, and T. Cukur. “COVID-19 Detection from Respiratory Sounds with Hierarchical Spectrogram Transformers.” *IEEE Journal of Biomedical and Health Informatics* 28, no. 3 (2024): 1273-1284.
2. He, W., Y. Yan, J. Ren, R. Bai, and X. Jiang. “Multi-View Spectrogram Transformer for Respiratory Sound Classification.” In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8626-8630, 2024.
3. Kim, B. J., J. H. Mun, D. H. Hwang, D. I. Suh, C. Lim, and K. Kim. “An explainable and accurate transformer-based deep learning model for wheeze classification utilizing real-world pediatric data.” *Scientific Reports* 15, article 5656 (2025).
4. Tawfik, M., and N. M. Al-Zidi. “Asthma Detection System: Machine and Deep Learning-Based Techniques.” In *Advances in Intelligent Systems and Computing*, vol. 1653, 205-218. Springer, Singapore, 2021. DOI: 10.1007/978-981-19-1653-3_16.
5. Zhantleuova, A. K., Y. K. Makashev, and N. T. Duzbayev. “Optimizing MFCC Parameters for Breathing Phase Detection.” *Sensors* 25, no. 16, article 5002 (2025).
6. Zhang, P., A. Swaminathan, and A. A. Uddin. “Pulmonary disease detection and classification in patient respiratory audio files using long short-term memory neural networks.” *Frontiers in Medicine* 10, article 1269784 (2023).
7. Tzeng, J.-T., J.-L. Li, H.-Y. Chen, Y.-C. Chen, W.-C. Cheng, S.-H. Wang, Y.-T. Lin, T.-W. Chang, and S.-H. Liu. “Improving the Robustness and Clinical Applicability of Automatic Respiratory Sound Classification Using Deep Learning–Based Audio Enhancement: Algorithm Development and Validation.” *JMIR AI* 4, article e67239 (2025).
8. Ifty, Tanzina Taher, et al. “Explainable Lung Disease Classification from Chest X-Ray Images Utilizing Deep Learning and XAI.” *arXiv preprint arXiv:2404.11428* (2024).
9. Coppock, Harry, et al. “Audio-based AI classifiers show no evidence of improved COVID-19 screening over simple symptoms checkers.” *Nature Machine Intelligence* 6, no. 2 (2024): 229-242.
10. Yan, Y., S. O. Simons, L. van Bemmel, P. H. N. de With, and F. van der Sande. “Optimizing MFCC parameters for the automatic detection of respiratory diseases.” *Applied Acoustics* 228, article 110299 (2025).
11. Xia, T., J. Han, and C. Mascolo. “Exploring machine learning for audio-based respiratory condition screening: A concise review of databases, methods, and open issues.” *Experimental Biology and Medicine* 247, no. 23 (2022): 2113-2133.

12. Ariyanti, W., K. C. Liu, K. Y. Chen, and T. Yu. "Abnormal Respiratory Sound Identification Using Audio-Spectrogram Vision Transformer." In *Proc. Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 1-4, 2023.
13. Bae, S., J.-W. Kim, W.-Y. Cho, D. Kim, S. Kim, and K. Jung. "Patch-Mix Contrastive Learning with Audio Spectrogram Transformer on Respiratory Sound Classification." In *Proc. INTERSPEECH*, 5436-5440, 2023.
14. Kwon, A. M., and K. Kang. "A temporal dependency feature in lower dimension for lung sound signal classification." *Scientific Reports* 12, article 7889 (2022).
15. Roy, A., B. Gyanchandani, A. Oza, M. Sahu, S. Jain, P. Sharma, and R. Wadhvani. "TriSpectraKAN: a novel approach for COPD detection via lung sound analysis." *Scientific Reports* 15, article 6296 (2025).
16. Xu, S., R. C. Deo, O. Faust, A. Rajendra Acharya, S. Chattopadhyay, and M. Haghi. "Automated Lightweight Model for Asthma Detection Using Respiratory and Cough Sound Signals." *Diagnostics* 15, no. 9, article 1155 (2025).
17. Wu, C., N. Ye, and J. Jiang. "Classification and Recognition of Lung Sounds Based on Improved Bi-ResNet Model." *IEEE Access* 12 (2024): 73079-73094.
18. Rishabh, D. Kumar, Y. Meena, and K. Singh. "Respiratory sound classification utilizing human auditory-based feature extraction." *Physica Scripta* 100, 046003 (2025).
19. Wanasinghe, T., et al. "Lung Sound Classification With Multi-Feature Integration Utilizing Lightweight CNN Model." *IEEE Access* 12 (2024): 21262-21276.
20. Srikar, D., K. M. Lakshmi, A. S. Kumar, I. Shivani, and S. Agarwal. "Respiratory disease classification using lung sounds with CNN-LSTM." *Journal of Computer Science* 18, no. 05 (2025): 243-249.
21. Ali, S. W., M. M. Rashid, M. U. Yousuf, S. Shams, M. Asif, M. Rehan, and I. D. Ujjan. "Towards the Development of the Clinical Decision Support System for the Identification of Respiration Diseases via Lung Sound Classification Using 1D-CNN." *Sensors* 24, 6887 (2024).
22. Hassan, U., A. Singhal, and P. Chaudhary. "Lung disease detection using EasyNet." *Biomedical Signal Processing and Control* 91, 105944 (2024).
23. Tzeng, J.-T., J.-L. Li, H.-Y. Chen, et al. "Improving the Robustness and Clinical Applicability of Automatic Respiratory Sound Classification Using Deep Learning-Based Audio Enhancement: Algorithm Development and Validation." *JMIR AI* 4, e67239 (2025).
24. Sreejith, R., et al. "Enhanced Lung Disease Classification Using CALMNet: A Hybrid CNN-LSTM-TimeDistributed Model for Respiratory Sound Analysis." *IEEE Access* 13 (2025): 135053-135073.
25. Nuha, M. F. F., B. D. M. S. P. Dassanayaka, H. M. S. N. Ariyadasa, M. Ramashini, and S. J. M. D. P. Samarakoon. "Optimization of lung sound classification by experimenting with different filtration techniques and LPCC feature extraction." *Sri Lankan Journal of Technology* 6, Special Issue (2025): 121-129.
26. Bhushan, P., et al. "A Self-Attention Based Hybrid CNN-LSTM Architecture for Respiratory Sound Classification." *GMSARN International Journal* 18 (2024): 54-61.
27. Khan, R., S. U. Khan, U. Saeed, and I.-S. Koo. "Auscultation-Based Pulmonary Disease Detection through Parallel Transformation and Deep Learning." *Bioengineering* 11, 586 (2024).
28. Roy, A., B. Gyanchandani, A. Oza, and A. Singh. "TriSpectraKAN: a novel approach for COPD detection via lung sound analysis." *Scientific Reports* 15, 6296 (2025).
29. Aljaddouh, B., M. Da, and F. Alaswad. "Multimodal Disease Detection and Classification Using Breath Sounds and Vision Transformer for Improved Diagnosis." *Procedia Computer Science* 235 (2024): 1436-1444.
30. Orkweha, K., K. Phapatanaburi, W. Pathonsuwan, T. Jumphoo, A. Rattanasak, P. Anchuen, W. Pinthurat, M. Uthansakul, and P. Uthansakul. "A Framework for Detecting Pulmonary Diseases from Lung Sound Signals Using a Hybrid Multi-Task Autoencoder-SVM Model." *Symmetry* 16, 1413 (2024).
31. Sabry, A. H., et al. "Machine learning approaches for respiratory disease detection: A comprehensive review." *Heliyon* 10, e26218 (2024).
32. Tawfik, M., Fathi, I. S., Nimbhore, S. S., Alsmadi, I. M., and Sawah, M. S."E-RespiNet: An LLM-ELECTRA driven triple-stream CNN with feature fusion for asthma classification." *PLOS ONE* 20, no. 11 (2025): e0334528. DOI: 10.1371/journal.pone.0334528.
33. Rocha, Bruno M., Dimitris Filos, Luis Mendes, Ioannis Vogiatzis, Eleni Perantoni, Evangelos Kaimakamis, Pantelis Natsiavas, et al. "An open access database for the evaluation of respiratory sound classification algorithms." *Physiological Measurement* 40, no. 3 (2019): 035001.
34. Fraiwan, Mohammad, Luay Fraiwan, Basheer Khassawneh, and Ali Ibnian. "A dataset of lung sounds recorded from the chest wall using an electronic stethoscope." *Data in Brief* 35 (2021): 106913.
35. De Cheveigné, Alain, and Hideki Kawahara. "YIN, a fundamental frequency estimator for speech and music." *The Journal of the Acoustical Society of America* 111, no. 4 (2002): 1917-1930.
36. Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks." *Proceedings of the International Conference on Machine Learning*. PMLR, 2017.
37. Kiran, Muhammad Saleem, Mesut Gündüz, and Önder Baykan. "A novel hybrid approach based on particle swarm optimization and ant colony algorithm to forecast energy demand of Turkey." *Energy Conversion and Management* 53, no. 1 (2012): 75-83.