Mühendislik Bilimleri ve Tasarım Dergisi
13(3), 896-910, 2025
e-ISSN: 1308-6693

Journal of Engineering Sciences and Design
DOI: 10.21923/jesd.1672180

Araştırma Makalesi Research Article

COUGH SOUND ANALYSIS WITH DEEP LEARNING: THE IMPACT OF DATA AUGMENTATION ON RESPIRATORY DISEASE CLASSIFICATION

Ayşen Özün TÜRKÇETİN[1,2*], Turgay KOÇ[3], Şule ÇİLEKAR[4]

[1] Graduate School of Natural and Applied Sciences, Mechanical Engineering, Suleyman Demirel University
[2] Career Planning and Alumni Communication Center, Suleyman Demirel University, Isparta, Türkiye
[3] Department of Electric Electronic Engineering, Suleyman Demirel University, Isparta,Türkiye
[4]Department of Pulmonology, Afyonkarahisar Health Sciences University, Afyonkarahisar, Türkiye

Keywords

Cough Sound Analysis, Lung Diseases, Deep Learning Models, Data Augmentation, Convolution Neural Network, Cross-Validation, Imbalanced Data.

Abstract

Respiratory diseases affect millions globally, necessitating efficient and early diagnostic tools to mitigate complications. This study proposes a robust and systematic approach for classifying asthma, COPD, pneumonia, and healthy conditions using cough sound analysis. Mel-frequency cepstral coefficients (MFCCs) were extracted and used to train both a deep learning model (CNN) and traditional classifiers (Random Forest, SVM) under limited and imbalanced data conditions. A major focus was on evaluating the impact of data augmentation and model choice on classification performance. Initial results showed that traditional models outperformed the CNN due to overfitting. However, with progressive augmentation up to 800 synthetic samples per class and the use of Dice Loss, the CNN model achieved substantial improvements, reaching 84% accuracy and a Macro F1 Score of 69%. These results highlight the critical role of data augmentation and tailored training strategies in enhancing the performance of deep learning models for audio-based biomedical classification tasks.

DERIN ÖĞRENME ILE ÖKSÜRÜK SESI ANALIZI: VERI ARTIRIMININ SOLUNUM YOLU HASTALIKLARI SINIFLANDIRMASI ÜZERINDEKI ETKISI

Anahtar Kelimeler

Öksürük Sesi Analizi, Akciğer Hastalıkları, Derin Öğrenme Modelleri, Veri Arttırma, Evrişimli Sinir Ağı, Çapraz Doğrulama, Dengesiz Veri.

Öz

Solunum yolu hastalıkları küresel olarak milyonlarca kişiyi etkileyerek komplikasyonları azaltmak için etkili ve erken tanı araçlarının gerekliliğini ortaya koymaktadır. Bu çalışma, öksürük sesi analizini kullanarak astım, KOAH, zatürre ve sağlıklı durumları sınıflandırmak için sağlam ve sistematik bir yaklaşım önermektedir. Mel-frekans cepstral katsayıları (MFCC'ler) çıkarılarak ve sınırlı olan dengesiz veri koşulları altında hem derin öğrenme modelini (CNN) hem de geleneksel sınıflandırıcıları (Rastgele Orman, SVM) eğitmek için kullanılmıştır. Çalışmanın başlıca odak noktası, veri artırmanın ve model seçiminin sınıflandırma performansı üzerindeki etkisini değerlendirmektir. İlk sonuçlar, aşırı uyum nedeniyle geleneksel modellerin CNN'den daha iyi performans gösterdiğini göstermiştir. Ancak, sınıf başına 800 sentetik örneğe kadar kademeli artırma ve Dice Loss kullanımıyla CNN modeli önemli iyileştirmeler elde ederek %84 doğruluk ve %69'luk bir Makro F1 Puanı elde edildi. Bu sonuçlar, ses tabanlı biyomedikal sınıflandırma görevleri için derin öğrenme modellerinin performansını artırmada veri artırmanın ve özel eğitim stratejilerinin kritik rolünü vurgulamaktadır.

Alıntı / Cite

Turkcetin A.O., Koc T., Cilekar S., (2025). Cough Sound Analysis with Deep Learning: The Impact of Data Augmentation on Respiratory Disease Classification, Mühendislik Bilimleri ve Tasarım Dergisi, 13(3), 896-910.

Yazar Kimliği / Author ID (ORCID Number)

Ayşen Özün Türkçetin, 0000-0003-4784-2267,
Turgay Koç, 0000-0002-4846-7772,
Şule Çilekar, 0000-0001-8659-955X

Makale Süreci / Article Process	
Başvuru Tarihi / Submission Date	08.03.2025
Revizyon Tarihi / Revision Date	06.07.2025
Kabul Tarihi / Accepted Date	30.07.2025
Yayım Tarihi / Published Date	30.09.2025

COUGH SOUND ANALYSIS WITH DEEP LEARNING:

* Corresponding author: aysenturkcetin@sdu.edu.tr, +90-246-211-8446

THE IMPACT OF DATA AUGMENTATION ON RESPIRATORY DISEASE CLASSIFICATION

Ayşen Özün Türkçetin[1,2†], Turgay Koç[3], Şule Çilekar[4],
[1]Graduate School of Natural and Applied Sciences, Suleyman Demirel University, Isparta, Türkiye
[2]Carrier and Planning Department, Suleyman Demirel University, Isparta, Türkiye
[3]Department of Electric Electronic Engineering, Suleyman Demirel University, Isparta,Türkiye
[4]Department of Pulmonology, Afyonkarahisar Health Sciences University, Afyonkarahisar, Türkiye

Highlights

- Respiratory disease classification was performed using a CNN model with noninvasive cough sound analysis.
- Following advanced data augmentation, the CNN model achieved an average F1-score of 69%, a macro AUC of 93%, and a mean Average Precision (mAP) of 74% across all respiratory disease classes, demonstrating its effectiveness in multi-class cough sound classification.
- The study demonstrates that audio-based data augmentation significantly improves model generalization and class balance in cough sound classification.

Purpose and Scope

The primary objective of this study is to develop a robust deep learning-based diagnostic tool for respiratory disease classification using cough sound analysis. This noninvasive approach aims to provide a scalable and cost-effective alternative to traditional diagnostic techniques. The study systematically evaluates the effect of data augmentation, class balancing, and feature extraction on model performance. By comparing CNN to various traditional machine learning models, the study contributes to AI-assisted healthcare solutions for respiratory diagnostics.

Design/methodology/approach

This study employs a deep learning-based framework for classifying respiratory diseases using cough sound recordings. The dataset comprises samples from individuals with Asthma, COPD, Pneumonia, and healthy controls. Four major audio augmentation techniques were applied to address class imbalance and expand the dataset. A CNN model was trained using Mel-frequency cepstral coefficients (MFCC), along with their delta and delta-delta features. A robust 3×3 Nested Cross-Validation scheme was used to ensure reliable evaluation under limited data conditions.

Findings

Applying multi-method data augmentation significantly improved classification performance. The CNN model achieved an average F1-score of 69%, a macro AUC of 0.95, and a mean Average Precision (mAP) of 0.74 across all respiratory disease classes. Without augmentation, traditional models like Random Forest performed better; however, with increased synthetic data, CNN outperformed all baselines. These results emphasize the importance of both data volume and targeted augmentation in achieving balanced and accurate predictions across all classes.

Research limitations/implications

The primary limitation of this study lies in the scarcity of real, labeled medical audio data, which is partially addressed via synthetic augmentation. Future research should explore real-time inference settings, alternative feature extraction techniques, and integration of other biosignals to improve system robustness.

Practical implications

This study presents a low-cost, noninvasive AI-assisted diagnostic approach that can help clinicians detect respiratory diseases earlier. Its integration into mobile applications or telemedicine platforms could reduce the burden on healthcare systems, particularly in resource-limited environments.

Social Implications

Using cough-based diagnostic systems can enhance healthcare accessibility in underprivileged regions. Early and automated detection of asthma, pneumonia, or COPD may help reduce mortality and long-term complications by enabling timely interventions.

Originality

Unlike many previous studies that focus on binary classification tasks such as distinguishing between healthy individuals and those with asthma or pneumonia, this study addresses the more complex challenge of simultaneously classifying three distinct respiratory diseases—Asthma, COPD, and Pneumonia—from cough sounds using a single deep learning framework. By applying multi-method data augmentation techniques, the proposed CNN model significantly improves classification performance across all classes. This comprehensive approach to multi-class cough-based diagnosis is rarely explored in the literature, making this work a unique contribution to the field of AI-based respiratory diagnostics.

† Corresponding author: aysenturkcetin@sdu.edu.tr, +90-246-211-8446

1. Introduction

Respiratory diseases affect millions of individuals globally, and early diagnosis of conditions such as asthma, chronic obstructive pulmonary disease (COPD), and pneumonia is critical for improving treatment outcomes and reducing long-term complications. Traditional diagnostic methods, though clinically reliable, are often expensive, time-consuming, and require expert intervention, limiting their accessibility—particularly in low-resource settings (World Health Organization, 2020; Brown et al., 2021).

To address these limitations, there has been growing interest in developing rapid, affordable, and noninvasive preliminary diagnostic tools using cough sounds—an easily collectible biosignal via commonly available devices like smartphones. Cough is a known biomarker that carries disease-specific acoustic characteristics, making it suitable for automated classification of respiratory conditions. Spectrogram and time-frequency analyses have revealed distinguishable patterns in cough sounds across different diseases, highlighting their diagnostic potential (Johnson et al., 2020; Smith et al., 2019).

Recent research demonstrates that artificial intelligence and machine learning techniques can classify conditions such as asthma, COPD, and pneumonia with high accuracy based on cough audio signals. For instance, Mel-Frequency Cepstral Coefficients (MFCC) and power spectral density features were used in one study, where a Support Vector Machine (SVM) with an RBF kernel achieved 95.86% accuracy in detecting COVID-19 from coughs (Melek, 2022). These findings confirm the effectiveness of advanced signal processing and classification methods for differentiating between respiratory conditions and suggest their applicability to broader respiratory disease detection.

Despite these promising advancements, challenges persist—particularly the limited availability of high-quality labeled medical audio data. Many existing datasets are small and imbalanced, often overrepresenting healthy individuals. Such data characteristics can lead to overfitting in deep learning models and hinder generalization across disease classes, thereby affecting diagnostic reliability.

This study proposes a comprehensive methodology for the automatic classification of respiratory conditions—including asthma, COPD, and pneumonia—through cough sound analysis. The approach integrates data augmentation, MFCC-based feature extraction, and classification models to enhance both accuracy and generalizability. Figure 1 illustrates the overall workflow.

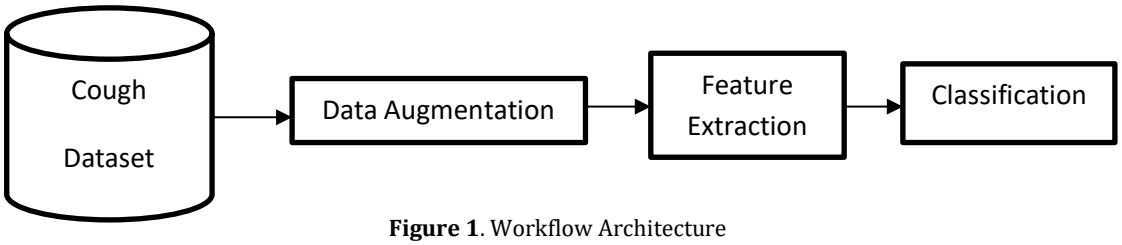


Figure 1. Workflow Architecture

In this work, a Convolutional Neural Network (CNN) model was trained on a four-class cough sound dataset (Asthma, COPD, Pneumonia, Healthy). The impact of different batch sizes and data augmentation levels on model performance was examined. Feature extraction relied on Mel-Frequency Cepstral Coefficients (MFCCs), which effectively capture the spectral properties of cough sounds.

The primary objective of this study is to systematically investigate how varying levels of data augmentation influence classification performance across both deep learning and traditional machine learning models, including Logistic Regression (LR), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Random Forest (RF). Unlike many existing studies that focus on a single model, this work aims to highlight how increased data size affects the comparative balance between CNN and traditional classifiers.

A robust nested cross-validation approach was employed to ensure unbiased model evaluation. Special emphasis was placed on metrics like Macro F1 Score and Mean Average Precision (mAP), which are more suitable for imbalanced datasets. The results show that data augmentation enhances model generalization and performance, particularly in deep learning models. These findings support the potential of cough sound analysis as an effective noninvasive diagnostic tool for respiratory diseases. Future research could explore the integration of more advanced augmentation strategies or hybrid model architectures to further improve classification performance

The remainder of this paper is organized as follows: Section 2 summarizes related work in the literature. Section

3 details the dataset used, feature extraction, data augmentation techniques, the classification models compared, and the experimental setup. Section 4 presents and analyzes the comparative results obtained under different scenarios. Finally, Section 5 discusses the findings and concludes the paper with the significance of the study and suggestions for future work.

2. Literature Survey

Respiratory diseases such as asthma, chronic obstructive pulmonary disease (COPD), and pneumonia affect millions of individuals worldwide, making early and accurate diagnosis essential. Traditional diagnostic methods, such as spirometry and radiographic imaging, are often invasive, expensive, and not always readily available. In recent years, advances in deep learning have enabled the development of non-invasive, audio-based diagnostic tools that analyze cough sounds for the detection of respiratory conditions. This section reviews key studies that have applied deep learning to cough and lung sound classification.

Data augmentation techniques have become an integral component of training robust deep learning models for biosignal classification. Schuller et al. (2020) examined the effects of various augmentation strategies—including band-pass filtering, loudness perturbation, additive noise, pitch shifting, and time stretching—on the classification of biomedical audio signals. Their results support the use of augmentation to improve generalization and resilience in deep learning applications for respiratory diagnostics.

Sharma et al. (2020) introduced the Coswara dataset and conducted a comparative analysis of CNNs, LSTMs, and hybrid models in the classification of respiratory conditions. Their results underline the high discriminative capacity of deep neural networks in cough-based diagnosis.

In their foundational survey, Pal and Sankarasubbu (2021) investigated deep learning approaches for cough-based COVID-19 diagnosis, emphasizing the importance of symptom embeddings and interpretable modeling. Their work highlights feature extraction strategies such as MFCC and spectrogram representations as critical components in audio-based diagnostic pipelines.

Pahar et al. (2021) employed transfer learning with pretrained models and bottleneck features for COVID-19 detection using cough, breath, and speech sounds. Using nested cross-validation, they demonstrated that deep learning models can be effectively adapted to respiratory audio classification even with limited labeled data.

Chakraborty et al. (2021) developed an AI-based cough classification system using smartphone recordings to detect chronic respiratory diseases like asthma, COPD, and pneumonia. Their pipeline included noise filtering, cough segmentation, MFCC feature extraction, and ensemble-based classification, achieving over 90% accuracy. However, their approach focused on binary or disease-specific models and did not examine the impact of data augmentation. In contrast, our study addresses these gaps through a CNN-based multi-class classification framework supported by systematic augmentation.

In a 2023 study, Celik et al. proposed CovidCoughNet, a CNN-based model incorporating pitch-shifting augmentation for COVID-19 detection. The model achieved 99.19% accuracy, though it focused solely on COVID-19 and did not address other respiratory diseases like asthma or COPD.

Vodnala et al. (2024) explored several deep learning models—including GRUs, LSTMs, and CNN-LSTM hybrids—to classify asthma and COPD from cough recordings. Their findings emphasize the potential of deep learning for differentiating between clinically overlapping respiratory conditions through cough analysis.

Shehab et al. (2024) developed a deep learning-based spectrogram classification system to detect a wide range of respiratory conditions, including pneumonia, pleural effusion, lung fibrosis, asthma, bronchitis, and COPD. Their use of spectrograms preserved critical temporal and frequency features, resulting in high diagnostic accuracy.

Sheikh et al. (2024) proposed a deep learning-based multilabel classification system for identifying multiple respiratory diseases from cough sounds, including asthma, bronchitis, pneumonia, and COPD. Their approach utilized mel spectrograms and CNN architectures, demonstrating the potential of multilabel models to detect overlapping conditions. However, their study was designed for multilabel prediction, where a single sample may belong to multiple classes simultaneously. In contrast, our study focuses on a multi-class classification setup—assigning one distinct class to each cough sample—which better reflects the practical clinical need for clear, mutually exclusive diagnoses. Furthermore, while Sheikh et al. did not systematically investigate data augmentation, our study incorporates multiple augmentation strategies and evaluates their effect on model performance under class imbalance.

In a multi-task learning approach, Suma et al. (2024) trained models to simultaneously classify lung sounds and predict related respiratory diseases. Their architecture included 2D CNNs, ResNet50, MobileNet, and DenseNet models, demonstrating that multi-task strategies can improve performance across multiple classification targets.

Finally, Alqudah and Moussavi (2025) provided a comprehensive review of deep learning techniques for biomedical signal processing. Their analysis emphasizes the role of architectural innovations (e.g., CNNs, RNNs, hybrid models) and advanced preprocessing methods in achieving high accuracy and robustness, reinforcing the validity of using MFCC and spectrogram-based input in respiratory disease classification tasks such as ours.

While previous studies have demonstrated the power of deep learning in classifying respiratory diseases, many have focused on a single model type or examined only a limited number of disease categories. Additionally, the use of simple train-test split methodologies in these works raises concerns about the generalizability of their findings. Beyond these methodological limitations, there are ongoing challenges related to feature selection, data augmentation strategies, and dataset imbalance. This study aims to address these gaps by conducting a systematic comparison between a deep learning model and multiple traditional machine learning algorithms across varying levels of data augmentation. Unlike prior research, which often concentrates on distinguishing a narrow range of illnesses, this study utilizes a larger and more diverse dataset that includes asthma, COPD, pneumonia, and healthy individuals.

To ensure the robustness of the results, a nested cross-validation approach was employed, offering a more reliable assessment than conventional validation methods. By expanding the dataset through augmentation and testing multiple architectures, this work overcomes the limitations of data scarcity and imbalance, especially for deep learning models that are prone to overfitting in such scenarios. The findings not only demonstrate the effectiveness of deep learning when supported by appropriate augmentation techniques but also contribute to the development of a practical, non-invasive diagnostic tool that can be used in real-world clinical applications.

3. Material and Method

3.1. Dataset

The original dataset used in this study consists of cough recordings collected from a total of 134 individuals. The data were obtained from patients diagnosed at the Department of Chest Diseases, Afyonkarahisar Health Sciences University and from publicly available data sources. The distribution of individuals and the number of 1000 ms cough segments (samples) obtained from them are as follows: 13 Asthma (86 samples), 28 COPD (174 samples), 14 Pneumonia (68 samples) and 79 Healthy individuals (222 samples). Ethical approval was obtained for all data and informed consent was obtained from the participants. This unbalanced structure of the dataset in terms of both the number of patients and the sample constitutes one of the main motivations of this study.

Cough sound recordings were primarily obtained from patients diagnosed with asthma, COPD, and pneumonia, alongside healthy individuals. In addition, 79 healthy cough samples from the publicly available COUGHVID dataset were integrated to enhance the dataset's representativeness for distinguishing between patients and healthy individuals. Each cough sound was recorded in a controlled clinical setting using a standardized protocol to minimize background noise and ensure consistency in data quality. This combined dataset serves as a valuable resource for developing machine learning models aimed at automated respiratory disease classification based on cough sound analysis.

3.2. Data Annotation

To prepare high-quality training data, a manual annotation process was performed on the audio recordings using a dedicated labeling tool (see Figure 2). In this interface, each respiratory event—including coughs, breathing sounds, and speech—was segmented and labeled based on both amplitude and waveform patterns. This manual approach allowed for precise temporal identification of target signals, which is especially critical for training deep learning models sensitive to temporal and spectral variations.

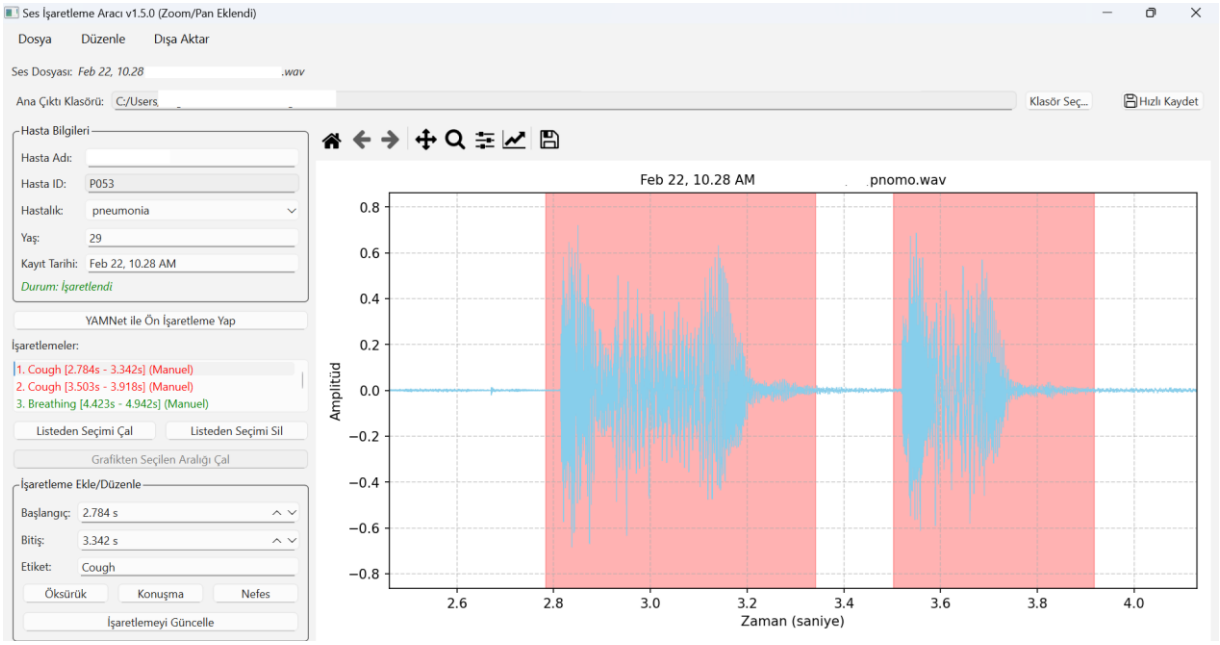


Figure 2. Manual annotation of cough and breathing segments using the sound labeling interface.

Each respiratory event was labeled manually to ensure accurate segmentation, which is essential for effective model training and performance evaluation. In the example shown, two cough segments and one breathing segment were manually marked in a pneumonia patient's audio sample. Segments were labeled based on auditory review and waveform analysis, ensuring accurate extraction of relevant sound events. Each segment was saved with metadata such as disease label, patient ID, and event type (e.g., cough or breathing), forming the basis for supervised model training. This annotation strategy ensured data consistency and improved the quality of the extracted features, ultimately enhancing classification performance.

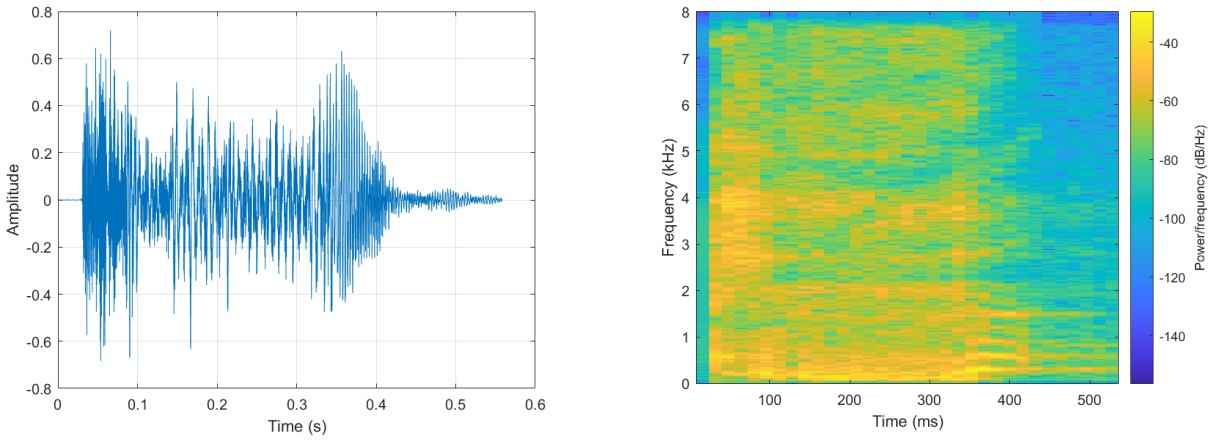


Figure 3. Spectrogram view of cough sound of an pneumatic patient with time and frequency bands

As shown in Figure 3, both the spectrogram and waveform of a pneumonia patient's cough provide valuable insights into the spectral energy distribution and temporal dynamics of the signal, which are essential for accurate classification using CNN-based models.

3.3. Data Augmentation Methodology

Audio data augmentation involves creating new training samples by applying transformations to original sounds, simulating real-world variations and improving model generalization by reducing overfitting. To increase the training data for the deep learning model and improve its generalization capability, this study implemented a comprehensive audio data augmentation pipeline. Four fundamental techniques were used: noise injection, pitch shifting, time shifting, and time stretching. Specifically, random Gaussian noise was added to the signal at 0.1–0.5% of its maximum amplitude; pitch was altered within −4 to +4 semitones; the signal was shifted up to 20% in time; and time was stretched between 0.8× and 1.2× without affecting pitch.

Using these methods, synthetic samples were generated for each class in the dataset (Healthy, Asthma, COPD, Pneumonia), addressing the challenge of class imbalance. Three experimental setups were created: one with no augmentation (baseline), one with 400 samples per class, and one with 800 samples per class in the training set. In the augmentation process, certain techniques such as noise injection were applied more frequently to underrepresented classes to achieve balance. The augmented dataset significantly improved diversity and robustness, enabling the model to learn more generalized features and reducing bias toward majority classes.

In order to balance the dataset, noise addition was performed more than once for some classes. In this way, each class of the dataset was increased by taking more samples. As a result of this process, the number of samples for each class was increased and the dataset was made more balanced.

The dataset initially contained varying numbers of original samples per class—221 for asthma, 196 for COPD, 98 for pneumonia, and 79 for healthy individuals—highlighting the class imbalance problem as shown in Table 1. Following the augmentation process, each class was expanded with thousands of synthetic samples generated using the described methods. This ensured a more balanced distribution of training data across all classes and enabled fairer comparisons among model performances during classification.

Table 1. Comparative table of cough sound samples according to diseases categories.

Diseases	Patient Count	Original Cough Sound Samples
Asthma	13	221
COPD	28	196
Pneumonia	14	79
Healthy	79	98

3.4. Proposed Method

One deep learning model and four traditional machine learning models were used for performance comparison. In this study, all models were used with their standard default parameters that are widely accepted in the literature. Audio data were processed to extract meaningful features for classification using a systematic approach. Cough sound recordings were obtained from patients diagnosed with asthma, COPD, pneumonia, and healthy individuals. The dataset was structured into training, validation, and test sets for model evaluation.

3.4.1 Data Preprocessing and Feature Extraction

Instead of using a traditional fixed train-validation-test split, this work used a 3×3 Nested Cross-Validation (NCV) strategy to achieve a robust and objective model evaluation. This methodical approach was chosen to eliminate the bias and unpredictability associated with random partitioning, especially given the limitations of the small and imbalanced dataset. The core principle of this strategy is to ensure a strict separation between the data used for model selection/tuning and the data used for final performance evaluation, thereby minimizing unintentional performance inflation and avoiding overfitting to a particular data split.
To implement this, all data partitions were performed at the patient level, guaranteeing that all samples from a single individual belong exclusively to either the training, validation, or test set within any given fold. The 3x3 NCV process consisted of two main loops:

1. The Outer Loop was responsible for final model evaluation. The patient-based dataset was partitioned into 3 folds. In each of the 3 iterations, two folds were used as the development set, while the remaining fold was held out as the final, unseen test set.
2. The Inner Loop handled model selection and hyperparameter tuning. For each development set provided by the outer loop, a separate 3-fold cross-validation was performed to find the optimal model configuration.

Crucially, data augmentation techniques were applied on-the-fly and exclusively to the training data within each inner loop. This ensured that both the validation sets (in the inner loop) and the final test sets (in the outer loop) remained in their original, unaltered state. This guarantees a truly unbiased evaluation of the model's ability to generalize to new, unseen patients. The final performance metrics reported in this study are the average of the results from the 3 outer loop test sets.

For the feature extraction procedure, Mel-Frequency Cepstral Coefficients (MFCCs) were employed, as they effectively capture the perceptual and spectral characteristics of sound—widely recognized in speech and bio-acoustic analysis. In addition to the standard 20 MFCCs, their first-order (delta) and second-order (delta-delta) derivatives were computed, resulting in a total of 60 features per audio chunk (20 MFCCs + 20 delta + 20 delta-delta).

The original 48 kHz audio recordings were downsampled to 16 kHz to ensure consistency and focus on relevant frequency bands for cough analysis. Each audio file was segmented into 1000 ms chunks to increase temporal resolution. Feature extraction was performed using the Librosa library, with the short-time Fourier transform (STFT) parameters set to a window size of 1024 and a hop length of 40. During preprocessing, all .wav files were iteratively loaded, features were transposed to match the model input format, and each segment was labeled based on its respective respiratory condition: Healthy, Asthma, COPD, or Pneumonia. Errors encountered during data loading were handled to preserve dataset integrity. These extracted features serve as the foundational input to the classification models, enabling them to learn both static and dynamic patterns within cough signals, and improving the accuracy of respiratory disease classification.

3.4.2 Traditional Machine Learning Models

For comparative analysis, several well-established traditional machine learning algorithms were implemented. These models were trained on the extracted MFCC features to evaluate their capability in classifying respiratory conditions based on cough sounds. The selected algorithms are commonly used in biomedical signal processing due to their interpretability, robustness, and computational efficiency.

Support Vector Machine (SVM): Mostly utilized for classification problems, SVM is a potent supervised learning technique. In a high-dimensional feature space, it finds the best hyperplane to maximally divide data points of various classes. SVM has demonstrated good generalization performance with less data and is especially useful when there are more characteristics than samples. SVM is flexible in a variety of fields, including medical diagnostics, since kernel functions may be used to accommodate non-linear boundaries (Cortes & Vapnik, 1995).

Random Forest: This ensemble learning method builds many decision trees during training and aggregates their results to reduce overfitting and increase prediction accuracy. A random subset of the data and features (bagging) is used to train each tree, and majority voting or averaging are usually used to arrive at final predictions. Random Forest has gained popularity in clinical research and health-related datasets because of its capacity to manage noisy and unbalanced data (Breiman, 2001).

K-Nearest Neighbors (KNN) is an instance-based learning technique that is straightforward but efficient. It assigns the most common class among the 'k' nearest training instances in the feature space in order to classify a new input. KNN is appropriate for complicated and irregular datasets since it makes no assumptions about the underlying distribution. Its classification performance in practical situations may be impacted by its sensitivity to feature scaling and data imbalance (Cover & Hart, 1967).

Logistic Regression: For applications involving binary and multi-class classification, logistic regression is a basic linear model. It uses the logistic (sigmoid) function to estimate the likelihood of a categorical dependent variable based on one or more independent factors. Despite its ease of use, Logistic Regression is a reliable baseline and is especially appreciated for its quick training time and interpretability. In medical data analysis, where model explainability and transparency are crucial, it is often utilized (Cox, 1958).

3.4.3 Deep Learning Models

To capture more complex and hierarchical patterns within the MFCC features, advanced deep learning architectures were designed. These models process the 2D MFCC spectrograms directly, leveraging their ability to learn abstract representations.

3.4.3.1 Convolutional neural network (CNN)

A typical 2D Convolutional Neural Network (CNN) is the main deep learning model used. Mel-frequency cepstral coefficient (MFCC) spectrograms of audio signals are one example of the grid-like data that this architecture is particularly made to analyze. As illustrated in Figure 4 the CNN model begins with an input layer that feeds into a series of convolutional blocks.

The first convolutional block consists of a Conv2D layer with 32 filters and a subsequent Dropout layer with a rate of 0.25. This is followed by a second convolutional block, which employs a Conv2D layer with 64 filters and another Dropout layer at 0.25 rate. A third convolutional block is then introduced, featuring a MaxPooling2D layer with a pool size of (2,2) to reduce dimensionality, followed by a Conv2D layer with 128 filters. All Conv2D layers utilize a (3,3) kernel size and 'same' padding to maintain feature map dimensions, ensuring the effective capture of local temporal and spectral patterns.

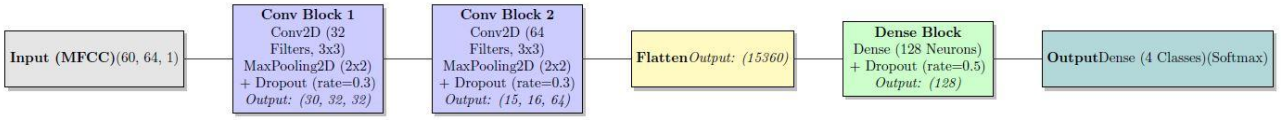


Figure 4. The architecture of the Convolutional Neural Network (CNN) model used in this study.

After two convolutional blocks and max-pooling, a Flatten layer prepares the feature maps for classification. The classification head consists of a 128-unit Dense layer with ReLU activation, followed by a Dropout layer (rate = 0.5) to reduce overfitting. The final output layer is a Dense layer with Softmax activation and four units, corresponding to the classification categories: Asthma, COPD, Pneumonia, and Healthy.

All deep learning models, including the hybrid architectures, incorporated Dropout layers to mitigate overfitting and enhance generalization capabilities. The final classification in each model is performed by a Dense layer with a Softmax activation function, yielding probabilities for the four distinct respiratory conditions. In the training phase, the Adam optimizer was employed alongside either the Categorical Cross-Entropy loss function or Dice loss (see. 3.5.4). The input data was segmented using two different chunk sizes (400 ms and 800 ms) to evaluate temporal effects. The learning rate was set to 0.0001, and the model was trained for 1000 epochs. To prevent overfitting, an early stopping mechanism was applied with a patience value of 10.

3.5 Performance Metrics

The confusion matrix served as the main indicator for assessing the machine learning model's efficacy. A table known as a confusion matrix is frequently used to describe the performance of a classification model, particularly with regard to its true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). For the computation of other important performance measures, including as accuracy, precision, recall, and F1-score, the confusion matrix is essential. The confusion matrix is crucial for computing other significant performance metrics, including as accuracy, precision, recall, and F1-score.

3.5.1 Accuracy: This is the most basic metric, representing the overall correctness of the model. It is defined as the ratio of correct predictions to total predictions:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

3.5.2 Precision: Precision gauges how well favorable forecasts turn out. It is calculated as the proportion of accurately anticipated positive observations to all projected positive observations.

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

3.5.3 Recall (Sensitivity): Recall gauges how well the model can identify every good example. It is the proportion of accurately forecasted favorable observations to all observations made during the actual class.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

3.5.4 The F1-Score and the Dice Score: The F1-Score and the Dice Score, also known as the Dice Similarity Coefficient or DSC, are mathematically equivalent metrics in binary classification tasks. Both metrics are harmonic

means that balance precision and recall, and are especially valuable when dealing with imbalanced datasets (Sørensen, 1948).

$$F1 - Score = Dice = 2\,X\,\frac{Precision*Recall}{Precision+Recall} \quad (4)$$

3.5.5 AUC-ROC: The True Positive Rate (TPR) and the False Positive Rate (FPR) are contrasted graphically in the ROC curve. The model's performance over a range of criteria is summarized by the AUC (Area Under the Curve), which measures the area under this curve. AUC values vary from 0 to 1, with 0.5 denoting random guessing and 1 denoting a flawless model.

Since these metrics offer a thorough grasp of the model's advantages and disadvantages, they were utilized to assess the model's performance across a range of classes. Studies like Chicco et al. (2020) and Berrar (2019) have emphasized the significance of these indicators in evaluating the performance of categorization algorithms.

4. Results and Analysis

The experimental results are presented in this section, providing a systematic analysis of model performance under different conditions. The findings begin with a baseline evaluation on the original dataset, followed by an in-depth examination of the effects of data augmentation, and conclude with the performance of the final, optimized model. The analysis integrates numerical data from the tables with visual evidence from the figures to offer a comprehensive understanding of the models' capabilities and stability.

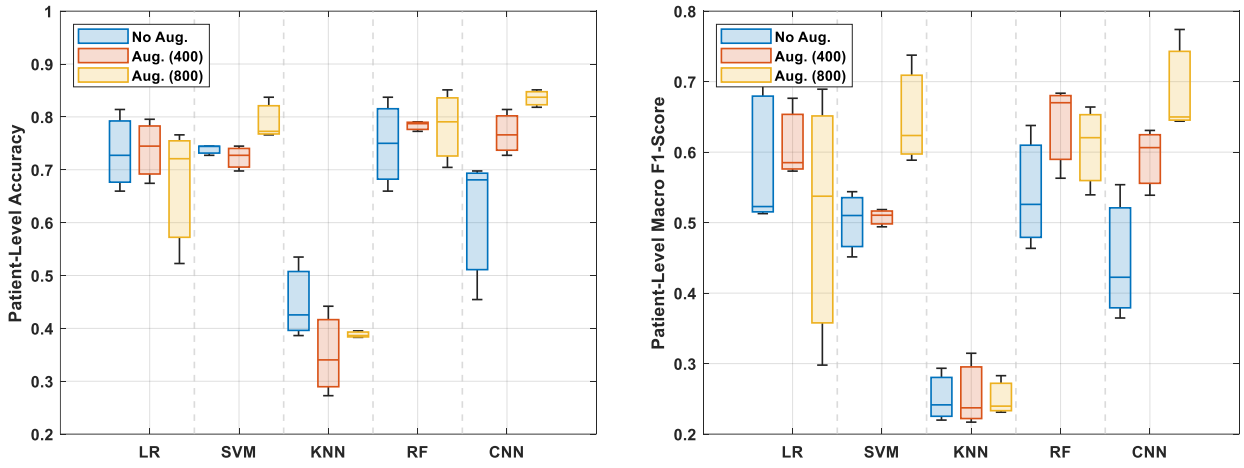
The initial experiments, conducted on the original dataset without any data augmentation, established a crucial performance baseline. The results of this baseline scenario are detailed through two complementary perspectives: Table 2 presents the mean performance values for the scenarios with no augmentation and with 800 augmented samples, while Figure 5 visualizes the full performance distribution across all three augmentation levels (0, 400, and 800 samples). In the limited data conditions without augmentation, traditional machine learning models demonstrated superior performance over the deep learning approach. Specifically, as shown in Table 2, Random Forest (RF) emerged as the top-performing model, achieving a Macro F1-Score of 54% and the highest mAP of 75%. Other traditional models like Logistic Regression (LR) and Support Vector Machine (SVM) also showed competent performance with F1-Scores of 58% and 50%, respectively, surpassing the CNN. The K-Nearest Neighbors (KNN) model, however, proved to be the weakest baseline with an F1-Score of only 25%. Beyond the mean scores, Figure 5 provides critical insight into model stability. The instability of the Convolutional Neural Network (CNN) was a key finding at this stage; its F1-Score of 44% was not only lower than most traditional models, but as its wide blue boxplot in Figure 5 illustrates, its performance exhibited high variance across the cross-validation folds, confirming its susceptibility to overfitting. In contrast, the narrow boxplot for RF indicates that it delivered a more consistent and stable performance.

Table 2. Comparative performance of all models across different data augmentation scenarios (mean values).

Model	F1-Score AUG OFF	F1-Score AUG 800	Accuracy AUG OFF	Accuracy AUG 800	mAP AUG OFF	mAP AUG 800	AUC AUG OFF	AUC AUG 800
LR	0.58	0.54	0.73	0.67	0.65	0.59	0.86	0.83
SVM	0.50	0.50	0.74	0.79	0.62	0.65	0.88	0.87
KNN	0.25	0.30	0.45	0.39	0.43	0.45	0.71	0.70
RF	0.54	0.54	0.75	0.78	0.75	0.74	0.91	0.91
CNN	0.44	0.60	0.61	0.77	0.64	0.74	0.94	0.95

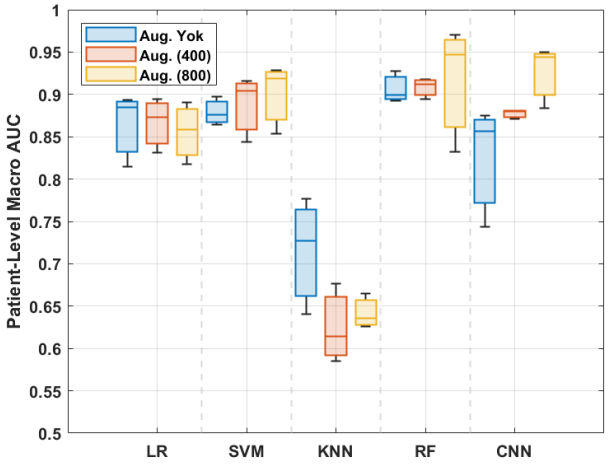
The introduction of data augmentation techniques served as a turning point in the comparative analysis, fundamentally altering the performance landscape. This process acted as a catalyst for the CNN model, addressing its initial weaknesses and ultimately establishing it as the most powerful model, while the performance of most traditional models remained largely stagnant. This evolution is evident through a combined analysis of the numerical data in Table 2 and the visual transformations in Figure 5. The CNN's journey from instability to robust performance began with a significant numerical leap; its Macro F1-Score increased from 44% to 60%, and its Macro AUC rose from 94% to 95%, demonstrating a marked improvement in its classification capability. This numerical improvement is visually substantiated by the metamorphosis of its boxplot in Figure 5. The journey from the blue (No Augmentation) to the yellow (800 Augmentation) state reveals two critical transformations: a significant upward movement on the vertical axis, confirming a substantial increase in median performance, and a visible narrowing of the box, proving that the variance in performance decreased and the model now produces

consistent and reliable results. In stark contrast, the traditional models could not leverage the synthetic data as effectively. The performance of Random Forest and SVM remained almost entirely unaffected, with their F1-Scores staying constant at 54% and 50%, respectively. Similarly, their boxplots in Figure 5 show little to no change across augmentation levels, confirming their limited capacity to learn from synthetic data. Other models like Logistic Regression (LR) even saw a slight decrease in performance, while K-Nearest Neighbors (KNN), the weakest baseline model, showed only a minor F1-Score improvement from 25% to 30%, which was insufficient to make it competitive. Consequently, data augmentation emerged as the key differentiator, allowing the CNN to overcome its initial instability and outperform all traditional classifiers in the 800-sample scenario, establishing its superiority in both raw performance metrics and stability.

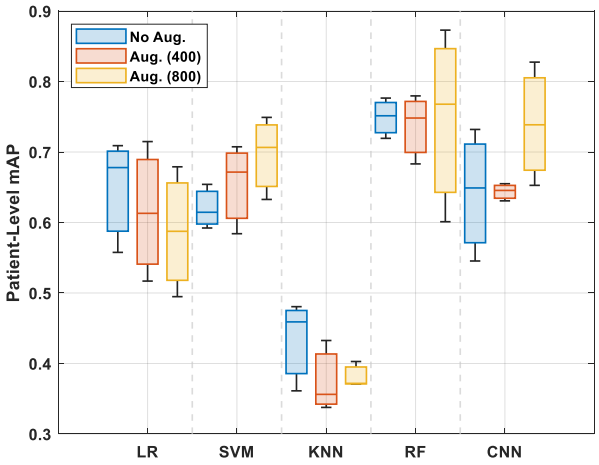


(a) Accuracy Distributions

(b) Macro F1-Score Distributions



(c) Macro AUC Distributions



(d) mAP Distributions

Figure 5. Performance distributions of all models at different data augmentation levels. Blue boxes represent the scenario with no augmentation, red boxes with 400 samples per class, and yellow boxes with 800 samples per class. The boxplots visualize model stability by showing the median (center line), interquartile range (the box), and overall spread (the whiskers) of the performance each model achieved during cross-validation.

To further enhance the performance of the CNN model, which had been stabilized through data augmentation, the Dice Loss function, sensitive to class imbalance, was employed. The impact of this final optimization is summarized in Table 3. The Dice Loss, which directly optimizes for the F1-Score metric, delivered a striking **14.26%** improvement, boosting the model's Macro F1-Score from 60% to 69%. This indicates that the model now recognizes minority classes much more successfully. Consequently, the overall patient-level accuracy also saw a significant rise to 84%, cementing the model's superior performance.

Table 3. Impact of Dice Loss on Augmented CNN Model Performance

Performance Metric	CNN (Cross-Entropy Loss)	CNN (Dice Loss)	Improvement
Patient Accuracy	0.77	0.84	+8.69%
Macro F1-Score	0.60	0.69	+14.26%
Macro AUC	0.90	0.93	+2.55%
mAP	0.75	0.74	-0.94%

Interestingly, the mAP (mean Average Precision) experienced a marginal decrease. This suggests a trade-off: while Dice Loss effectively improves the classification of underrepresented classes (boosting recall and F1-score), it may slightly alter the precision-recall curve for dominant classes.

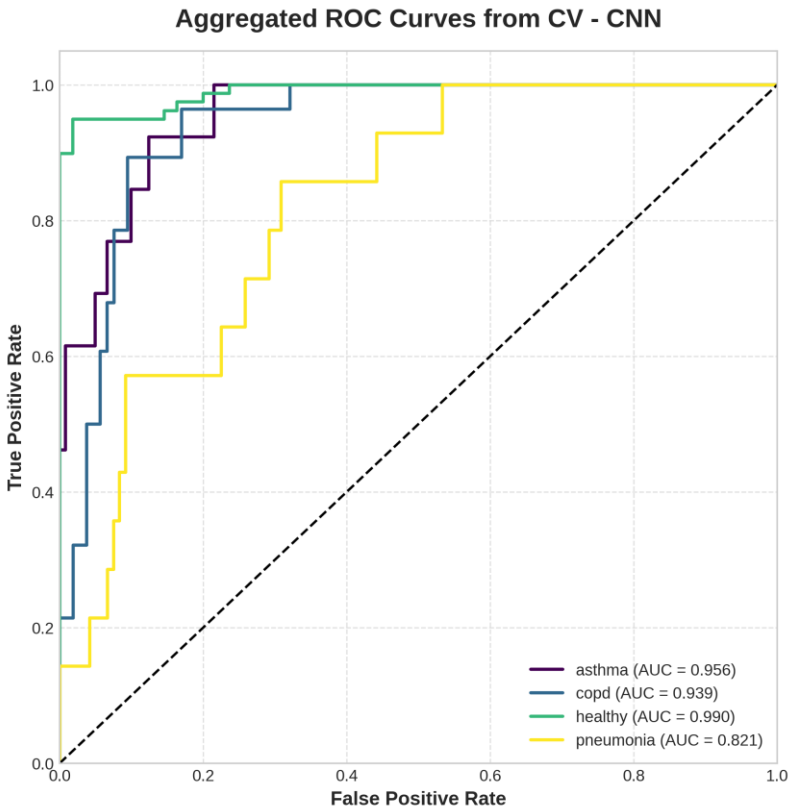


Figure 6. ROC curve analysis showed that the model was effective in the classification of respiratory diseases.

The class-specific discriminative power of the final, fully optimized CNN model is demonstrated by the ROC curve analysis shown in Figure 6. The model exhibited near-perfect performance for classes like Healthy (AUC: 0.99) and COPD (AUC: 0.96). Most importantly, it achieved a strong AUC of 0.90 for Asthma and 0.82 for Pneumonia, the most challenging and underrepresented class in the original dataset. This result proves that the applied two-stage methodology not only increased overall performance but also created a balanced and robust model capable of overcoming the difficulties posed by imbalanced medical data.

5. Discussion

This section interprets the results presented in Section 4, contextualizes them within the existing literature, and discusses the implications and limitations of the study.

5.1. Interpretation of Key Findings

The results reveal a clear, two-step path to developing a high-performance cough-based diagnostic tool under challenging data conditions. The initial experiments confirmed that on small, imbalanced datasets, traditional models like Random Forest can outperform deep learning models due to the latter's tendency to overfit. However, the two key strategies employed data augmentation and the use of a class-aware loss function (Dice Loss) worked in synergy to unlock the CNN's potential. Data augmentation provided the necessary feature diversity for the model to learn from, while Dice Loss guided the model to pay sufficient attention to minority classes during

training. This two-pronged approach was directly responsible for the substantial increase in the model's F1-Score and overall accuracy, establishing the CNN as the superior model for this task. The visual analysis of performance distributions further confirmed that these strategies not only improved the average performance but also increased the model's stability and consistency.

5.2. Comparison with the Literature

A direct numerical comparison of performance metrics with previous studies is challenging due to significant differences in datasets, class definitions, and evaluation protocols. However, it is valuable to contextualize our findings within the broader body of research. The performance of our final CNN model, achieving a Macro AUC of 0.93 and a Macro F1-Score of 0.69, falls within the competitive range reported for respiratory sound classification. More importantly, the primary contribution of our work lies in its methodological distinction. Unlike many previous studies that focused on binary tasks, our framework successfully addresses the more complex challenge of simultaneously classifying four distinct classes. Furthermore, our systematic investigation into the effects of data augmentation, validated with a rigorous nested cross-validation scheme, addresses a common gap in the literature where augmentation is often used without a comparative analysis of its impact.

5.3. Implications of the Study

The findings have significant practical and social implications. This study presents a low-cost, noninvasive AI-assisted diagnostic approach that can be integrated into mobile applications or telemedicine platforms. Such a tool could help clinicians detect respiratory diseases like asthma, COPD, and pneumonia earlier, which is critical for improving patient outcomes. Particularly in resource-limited environments where traditional diagnostic tools are scarce, a cough-based system can enhance healthcare accessibility, reduce the burden on healthcare systems, and potentially lower mortality rates by enabling timely interventions.

5.4. Limitations and Future Work

Although the findings of this study are promising, some limitations should be considered to contextualize the results. First**,** the performance of the developed models has not been validated on a fully independent, external test dataset. The primary reason for this is the current lack of a publicly available cough dataset that perfectly matches the multi-class scope of this study (Asthma, COPD, Pneumonia, and Healthy). Second**,** a large portion of the dataset was collected from a single clinical center, which may limit the generalizability of the models to different populations and recording environments.

Future work will focus on addressing these limitations and exploring broader research avenues. Our immediate plans involve (1) adapting our model for external validation on other public datasets, even if it requires testing on subsets of diseases, and (2) expanding our data collection to include multiple centers.
Beyond these steps, further research could significantly advance the field. A critical long-term goal is the creation of a large-scale, publicly available "benchmark" dataset through the joint efforts of researchers from different hospitals and geographical regions. Such a dataset is indispensable for enabling reproducible research and providing a solid foundation for proving the generalizability of new models. Furthermore, we propose exploring more advanced techniques, such as enriching datasets using advanced synthetic data generation methods, like Generative Adversarial Networks (GANs) and examining the potential of Transformer-based models to capture long-term temporal dependencies in cough sounds more effectively than traditional CNNs.

Pursuing these directions would be valuable for developing more robust and accurate AI-driven diagnostic tools for respiratory health.

Acknowledgement

In this study, the dataset was collected from patients hospitalized in the Department of Chest Diseases, Afyonkarahisar Health Sciences University. This study is a part of Ayşen Özün Türkçetin's doctoral dissertation. We thank Afyonkarahisar Health Sciences University for her help during the ethics committee and dataset stages.

Conflict of Interest

No conflict of interest was declared by the authors.

References

Allamy, S., & Koerich, A. L. (2021). 1D CNN architectures for music genre classification. In 2021 IEEE symposium series on computational intelligence (SSCI) (pp. 01-07). IEEE.

Alqudah, A. M., & Moussavi, Z. (2025). A Review of Deep Learning for Biomedical Signals: Current Applications, Advancements, Future Prospects, Interpretation, and Challenges. Computers, Materials & Continua, 83(3), 3021-3047.

Balamurali, B. T., Hee, H. I., Kapoor, S., Teoh, O. H., Teng, S. S., Lee, K. P., ... & Chen, J. M. (2021). Deep neural network-based respiratory pathology classification using cough sounds. Sensors, 21(16), 5555.

Berrar, D. (2019). "Accuracy and Precision: Evaluating the Performance of Machine Learning Models." *Data Science Journal*, 18(3), 102-113.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32.

Brown, C., Nissen, I., & Smith, R. (2021). Deep learning applications in biosignal analysis: A review of noninvasive diagnostics. Journal of Medical AI, 8(2), 112-130.

Celik, G. (2023). CovidCoughNet: A new method based on convolutional neural networks and deep feature extraction using pitch-shifting data augmentation for covid-19 detection from cough, breath, and voice signals. Computers in Biology and Medicine, 163, 107153.

Chakraborty, S., Ghosh, P., Bhattacharya, M., Dutta, S., Banerjee, A., & Sinha, R. (2021). An AI-based cough recognition and classification system using smartphone audio recordings for early diagnosis of chronic diseases. PLOS ONE, 16(11), e0259021. https://doi.org/10.1371/journal.pone.0259021.

Chicco, D., et al. (2020). "A Comprehensive Review on Performance Metrics for Classification Models." *Journal of Machine Learning Research*, 21(1), 1-45.

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21–27.

Cox, D. R. (1958). The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B, 20(2), 215–242.

Dey, R., & Salem, F. M. (2017). Gate-variants of gated recurrent unit (GRU) neural networks. In 2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS) (pp. 1597-1600). IEEE.

Farzad, A., Mashayekhi, H., & Hassanpour, H. (2019). A comparative performance analysis of different activation functions in LSTM networks for classification. Neural Computing and Applications, 31, 2507-2521.

Graves, A., Jaitly, N., & Mohamed, A. R. (2013). Hybrid speech recognition with deep bidirectional LSTM. In 2013 IEEE workshop on automatic speech recognition and understanding (pp. 273-278). IEEE.

Hochreiter, S. (1997). Long Short-term Memory. *Neural Computation MIT-Press*.

Johnson, M., et al. (2020). "Analysis of cough sound features for diagnosing respiratory conditions." Journal of Medical Acoustics, 12(3), 145-158.

Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and Music Signal Analysis in Python. Proceedings of the 14th Python in Science Conference, 18–25.

Melek Manshouri, N. (2022). Identifying COVID-19 by using spectral analysis of cough recordings: a distinctive classification study. Cognitive neurodynamics, 16(1), 239-253.

Milletari, F., Navab, N., & Ahmadi, S. A. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. In 2016 Fourth International Conference on 3D Vision (3DV) (pp. 565–571). IEEE. https://doi.org/10.1109/3DV.2016.79.

Pahar, M., O'Connell, O., et al. (2021). COVID-19 detection in cough, breath and speech using deep transfer learning and bottleneck features. npj Digital Medicine, 4(1), 166.

Pal, A., & Sankarasubbu, M. (2021). Pay attention to the cough: Early diagnosis of COVID-19 using interpretable symptoms embeddings with cough sound signal processing. In Proceedings of the 36th Annual ACM Symposium on Applied Computing (pp. 620-628).

Schuller, B., Batliner, A., Steidl, S., & O'Reilly, J. (2020). Data augmentation strategies for improving biosignal classification. IEEE Transactions on Biomedical Engineering, 67(5), 1450-1462.

Sharma, N., Krishnan, P., Kumar, R., Ramoji, S., Chetupalli, S. R., Ghosh, P. K., & Ganapathy, S. (2020). Coswara--a database of breathing, cough, and voice sounds for COVID-19 diagnosis. arXiv preprint arXiv:2005.10548.

Shehab, S. A., Mohammed, K. K., Darwish, A., & Hassanien, A. E. (2024). Deep learning and feature fusion-based lung sound recognition model to diagnoses the respiratory diseases. Soft Computing, 1-17.

Sheikh, K. A., Patel, B., Shah, R., & Shah, M. (2024). Deep learning-based multilabel classification of cough sounds for screening of respiratory diseases. PLOS ONE, 19(2), e0289317. https://doi.org/10.1371/journal.pone.0289317.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), 1-48.

Smith, J., et al. (2019). "Cough sound analysis in chronic obstructive pulmonary disease." Respiratory Medicine, 75(5), 230-240.

Sørensen, T. J. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. Kongelige Danske Videnskabernes Selskab, 5(4), 1–34.

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., & Jorge Cardoso, M. (2017). Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3 (pp. 240-248). Springer International Publishing.

Suma, K. V., Koppad, D., Kumar, P., Kantikar, N. A., & Ramesh, S. (2024). Multi-task Learning for Lung Sound and Lung Disease Classification. SN Computer Science, 6(1), 51.

Vodnala, N., Yarlagadda, P. S., Ch, M., & Sailaja, K. (2024). Novel Deep Learning Approaches to Differentiate Asthma and COPD Based on Cough Sounds. In 2024 Parul International Conference on Engineering and Technology (PICET) (pp. 1-4). IEEE.

World Health Organization. (2020). Global burden of respiratory diseases and diagnostic challenges. WHO Reports, 15(3), 45-60.

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. Advances in neural information processing systems, 28.

Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., & Xu, B. (2016). Attention-based bidirectional long short-term memory networks for relation classification. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 207-212.