

VoxMed: One-Step Respiratory Disease Classifier using Digital Stethoscope Sounds

*Paridhi Mundra**, *Manik Sharma**, *Yashwardhan Chaudhuri**, *Orchid Chetia Phukan*, *Arun Balaji Buduru*

IIT-Delhi, India
*equal contribution

{paridhi20392, yashwardhan20417, manik21336, orchidp, arunb}@iiitd.ac.in

Abstract

As respiratory illnesses become more common, it's crucial to quickly and accurately detect them to improve patient care. There is a need for improved diagnostic methods for immediate medical assessments for optimal patient outcomes. This paper introduces VoxMed—a UI-assisted one-step classifier that uses digital stethoscope recordings to diagnose respiratory diseases. It employs an Audio Spectrogram Transformer (AST) for feature extraction and a 1-D CNN-based architecture to classify respiratory disease, offering professionals information regarding their patients' respiratory health in seconds. We use the ICBHI dataset, which includes stethoscope recordings collected from patients in Greece and Portugal, to classify respiratory diseases. [github repository: https://github.com/Sample-User131001/VoxMed](https://github.com/Sample-User131001/VoxMed)

Index Terms: speech recognition, Disease Detection,

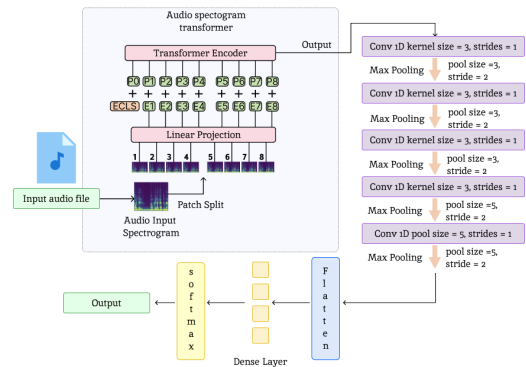


Figure 1: **VoxMed Architecture:** The architecture accepts a wav file for digital stethoscope sound and passes through an audio spectrogram transformer to extract features. Extracted features are passed through a 1-D CNN architecture as shown in the image to detect the type of respiratory disease.

1. Introduction

Respiratory-related diseases, such as chronic obstructive pulmonary disease and asthma, are leading causes of mortality globally [1], making a swift and dependable respiratory diagnosis important for averting any respiratory complication.

We introduce VoxMed, a simple UI system that detects respiratory diseases using digital stethoscope sounds. Our goal is to overcome the challenges of traditional respiratory diagnostics, which typically involve long waits in crowded clinical settings[2]. With VoxMed, healthcare practitioners can efficiently respond to patients’ medical needs based on fast respiratory health assessments. Past studies[3] have shown that machine learning algorithms can potentially evaluate medical audio data to identify illnesses[4]. VoxMed elaborates on the previous works to develop a one-step interface that enables doctors to record patient stethoscope sounds and support seamless diagnosis and analysis of patients. VoxMed accomplishes this by employing an Audio Spectrogram Transformer(AST)[5] to extract features and a 1-D Convolutional Neural Network (CNN) architecture for classifying diseases, as shown in Figure 1. We test VoxMed’s performance using the ICBHI challenge dataset[6], featuring stethoscope sounds gathered from patients in Greece and Portugal. Our findings demonstrate that VoxMed reliably identifies respiratory diseases such as COPD, potentially enhancing patient care. Our findings illustrate the platform’s ability to reliably identify respiratory disorders, underlining its potential as a useful tool for respiratory disease diagnosis and therapy in clinical practice.

2. Application Overview

Our work uses a simple 1-step respiratory disease detection process, illustrated in Figure 2, where users submit digital stethoscope sound samples of any duration. The model internally processes the audio using AST to make audio embeddings that are passed to a 1D-CNN, for identifying potential patient diseases. Our current application categorizes diseases as Healthy, COPD, or other potential diseases like URTI, LRTI, and Bronchitis.

Data Details: We utilize the ICBHI Sound Database[6] for our implementation, comprising 920 annotated recordings from 126 patients, totaling around 6898 respiratory cycles. The database provides comments for crackles, wheezes, and combinations, offering both clean and noisy respiratory recordings reflecting real-life scenarios and featuring patients of diverse ages.

Feature Extraction and Model Architecture: Our model utilizes AST for audio embedding generation, operating at a 16000 sampling rate. Figure 1 showcases the CNN architecture tailored for classification tasks, featuring multiple layers like convolutional, max-pooling, and fully connected layers. The input layer employs a 1D convolutional layer with 256 filters and a size of 3, followed by ReLU activation. Subsequent layers include decreasing filter sizes and max-pooling, with dropout regularization applied to prevent overfitting. The final layer consists of a softmax activation function with six units corresponding to classification classes.

Model Compilation and Training: The model is built with the Adam optimizer and a categorical cross-entropy loss function. we use an 80-20 split to train and evaluate our model.

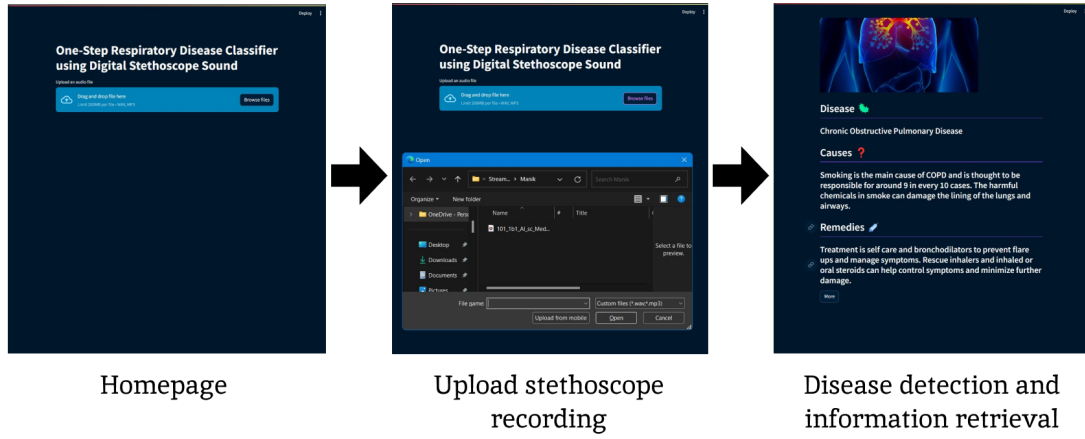


Figure 2: **VoxMed Workflow:** VoxMed respiratory detection requires the input of a digital stethoscope recording from the patient. We upload the recording through the UI and click on submit to process. Finally, we see possible respiratory ailment along with its symptoms and more information gathered through scraping information using APIs

Models	Accuracy	F1
Healthy/COPD/others		
AST(VoxMed)[5]	0.90	0.70
Wav2Vec2[7]	0.88	0.48
Unispeech[8]	0.89	0.58
WavLM[9]	0.86	0.47
Healthy/COPD/URTI/others		
AST(VoxMed)[5]	0.85	0.69
Wav2Vec2[7]	0.90	0.67
Unispeech[8]	0.86	0.43
WavLM[9]	0.86	0.44
Healthy/COPD/URTI/LRTI/Bronchitis		
AST(VoxMed)[5]	0.90	0.67
Wav2Vec2[7]	0.90	0.48
Unispeech[8]	0.89	0.54
WavLM[9]	0.89	0.53

Table 1: Ablations: Class-wise performance of VoxMed With different feature extraction backbones; F1 is macro-average F1 socre

3. Evaluation

Ablations: We assess VoxMed’s performance on a class-by-class basis, utilizing macro weighted F1 score and accuracy metrics, as detailed in Table 1. Experimentation involved testing various feature extraction networks within VoxMed, including Wav2Vec2[7], Unispeech[8], AST[5], and WavLM[9]. Our analysis reveals that AST consistently outperforms the other methods across both Macro F1 and accuracy measurements. Notably, the classification criterion of Healthy/COPD/others demonstrates the highest F1 score, suggesting heightened reliability in disease detection using this specific approach.

4. Conclusion

VoxMed emerges as a solution for rapid respiratory disease diagnosis, offering healthcare professionals immediate insights

into patient conditions. By leveraging an AST feature extraction and a 1-D CNN-based classifier, VoxMed achieves competitive performance along with a seamless 1-step UI. Its ability to streamline diagnostic processes and provide timely medical assessments underscores its potential to significantly impact patient care and healthcare delivery.

5. References

- [1] V. Cukic, V. Lovre, D. Dragisic, and A. Ustamujic, “Asthma and chronic obstructive pulmonary disease (copd)—differences and similarities,” *Materia socio-medica*, vol. 24, no. 2, p. 100, 2012.
- [2] K. R. Brekke, L. Siciliani, and O. R. Straume, “Competition and waiting times in hospital markets,” *Journal of Public Economics*, vol. 92, no. 7, pp. 1607–1628, 2008.
- [3] E.-A. Paraschiv and C.-M. Rotaru, “Machine learning approaches based on wearable devices for respiratory diseases diagnosis,” in *2020 International Conference on e-Health and Bioengineering (EHB)*. IEEE, 2020, pp. 1–4.
- [4] S. Gairola, F. Tom, N. Kwatra, and M. Jain, “Respirenet: A deep neural network for accurately detecting abnormal lung sounds in limited data setting,” in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 527–530.
- [5] Y. Gong, Y.-A. Chung, and J. Glass, “AST: Audio Spectrogram Transformer,” in *Proc. Interspeech 2021*, 2021, pp. 571–575.
- [6] B. Rocha, D. Filos, L. Mendes, I. Vogiatzis, E. Perantoni, E. Kaimakamis, P. Natsiavas, A. Oliveira, C. Jácome, A. Marques *et al.*, “A respiratory sound database for the development of automated classification,” in *Precision Medicine Powered by pHealth and Connected Health: ICBHI 2017, Thessaloniki, Greece, 18-21 November 2017*. Springer, 2018, pp. 33–37.
- [7] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
- [8] C. Wang, Y. Wu, Y. Qian, K. Kumtani, S. Liu, F. Wei, M. Zeng, and X. Huang, “Unispeech: Unified speech representation learning with labeled and unlabeled data,” in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 937–10 947.
- [9] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.