Non-Invasive TB Detection using Acoustic and Semantic Features from Cough Sounds

Yasmeena Akhter, Rishabh Ranjan, Bikash Dutta, Mayank Vatsa and Richa Singh✉

Indian Institute of Technology, Jodhpur, India
{akhter.1, ranjan.4, d22cs051, mvatsa, richa}@iitj.ac.in

Abstract. We present a novel dual-stream deep learning architecture, *AcouSem-AFNet*, for automated tuberculosis (TB) detection using acoustic analysis of respiratory sounds. The proposed architecture utilizes two complementary pathways to extract distinct semantic and acoustic characteristics essential for identifying TB-related respiratory patterns. Specifically, the semantic stream employs a Whisper encoder to model structured patterns in respiratory events, while the acoustic stream leverages WavLM to capture detailed temporal dynamics characteristic of TB cough sounds. These distinct features are fused through a specialized backbone with squeeze-excitation mechanisms and residual connections, designed explicitly to maintain discriminative capabilities and mitigate overfitting challenges typical of limited medical datasets. Evaluated on the CODA-TB challenge dataset, our approach achieves state-of-the-art performance with an accuracy of 78.10% and an AUC of 0.79, demonstrating improvements of 3% in AUC and 2% in accuracy over leading baseline methods. Our framework enables rapid, non-invasive TB screening, particularly beneficial for resource-limited settings, demonstrating the feasibility of deep learning-based acoustic analysis as a scalable, preliminary diagnostic tool to enhance global TB screening accessibility. The code and models are publicly available at `https://github.com/IAB-IITJ/AcouSem-AFNet`.

Keywords: Tuberculosis · CAD · Cough · Acoustic Features

1 Introduction

Tuberculosis (TB), caused by Mycobacterium tuberculosis, remains a critical global health challenge, with 9.9 million cases and 1.3 million deaths reported in 2020 [1]. TB, one of humanity's oldest infectious diseases [2], has claimed an estimated one billion lives throughout history [3]. Currently ranking as the leading cause of death among infectious diseases globally, TB continues to pose a significant public health challenge with approximately 10 million new active cases reported annually [4]. Despite its severity, approximately 40% of TB cases go undiagnosed due to limited healthcare access and diagnostic barriers [5]. This significant gap in detection highlights the urgent need for accessible, non-invasive screening methods.

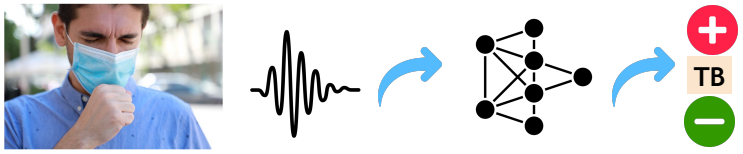

Fig. 1: Highlights the visual abstract, addressing the AI-based automatic TB detection from the cough sounds.

Traditional TB screening through self-reported symptoms presents critical limitations: low specificity leading to over-testing, and stigma-induced under-reporting resulting in inadequate care [6, 7]. This necessitates the development of objective, low-cost point-of-care screening methods. Acoustic-based TB detection emerges as a promising specimen-free, automated screening solution that could optimize resource allocation for molecular testing [8].

Cough, being a primary symptom of TB, presents a promising acoustic biomarker for automated disease detection [9]. The scientific basis for acoustic TB detection lies in the pathophysiology of respiratory diseases. Different pathological conditions alter airway dynamics and glottal behaviour, producing distinctive acoustic signatures during cough episodes [10]. Previous research has demonstrated the feasibility of distinguishing various respiratory conditions (asthma, bronchitis, pertussis) through cough acoustics, suggesting the viability of automated TB detection through acoustic pattern analysis [11, 12], as also demonstrated in Figure 1 as the visual abstract for this problem statement.

Recent advances in deep learning and acoustic signal processing have enabled the development of sophisticated audio analysis systems. AI models offer a transformative approach to respiratory illness diagnosis by providing non-invasive, cost-effective, and real-time data interpretation capabilities, addressing the limitations of traditional diagnostic methods and presenting particular value for resource-constrained settings. While preliminary studies [13, 14]have demonstrated the potential of cough-based TB screening, these efforts were limited by small sample sizes and restricted settings. Our work addresses these limitations by introducing a novel deep-learning framework, *AcouSem-AFNet*, that leverages raw audio-based features for robust TB classification from cough sounds. This research advances the field in three key aspects: (1) we propose a new architecture optimized for acoustic TB biomarker detection, (2) we demonstrate state-of-the-art performance on the comprehensive CODA-TB dataset, and (3) we provide an extensive comparison with different models.

2 Related Work

Respiratory disease detection through cough analysis has shown promising results across multiple conditions. Using cough frequency analysis, Marsden et al. [15] successfully detected Asthma Bronchiale (AB). Infante et al. [16] demonstrated cough sounds for screening AB, COPD, and TB. Windmon et al. [17] employed random forest classifiers for early CHF and COPD detection. Recent advances include acoustic feature extraction (MFCCs, ZCR) for pneumonia [18]

and COVID-19 classification [14, 19], while [20] demonstrated the effectiveness of generative adversarial networks for respiratory disease classification.

In the existing literature on TB detection through cough sounds, Tracey et al. [21] and Larson et al. [22] linked cough frequency to TB recovery. Pahar et al. [14] achieved high specificity (95%) and sensitivity (93%) using logistic regression on a small dataset (16 TB patients, 35 healthy controls). In follow-up work, Pahar et al. [19] tested multiple deep learning models on 47 TB patients, with ResNet50 [23] achieving 92.59% accuracy in TB vs. COVID-19 classification and 86.31% in three-class discrimination. Frost et al. [24] implemented Bi-LSTM on 74 subjects (28 TB patients, 46 controls), achieving 75% specificity and 89% sensitivity. These studies demonstrate that cough acoustics contain significant discriminative information for the classification of respiratory diseases, suggesting a potential for developing automated screening tools.

3 Methodology

The proposed methodology, *AcouSem-AFNet*, for TB detection from respiratory audio recordings employs a dual-stream architecture integrated with a RawNet3-based backbone [25] network specifically optimized for medical audio analysis. The system processes raw audio inputs through parallel semantic and acoustic pathways to extract complementary features crucial for identifying TB-related respiratory patterns, including distinctive cough characteristics, breathing anomalies, and associated acoustic signatures. The semantic stream, powered by the Whisper-large encoder, captures structured patterns in respiratory events, while the acoustic stream, utilizing WavLM [26], preserves fine-grained temporal characteristics specific to TB-related sounds. These complementary features are integrated through a specialized backbone network incorporating squeeze-excitation mechanisms and residual connections, designed to maintain discriminative power while preventing overfitting on limited medical datasets. The complete processing pipeline consists of two main components: (1) a front-end feature extraction module utilizing dual-stream processing for comprehensive respiratory sound analysis and (2) a backend network featuring AFMS-Res2MP blocks with targeted regularization for robust TB detection.

3.1 Audio Encoders and Modality Adapters

We propose a dual-stream design that processes semantic and acoustic cues in parallel to capture acoustic features for TB detection from audio cough recordings. Several works [26, 27] demonstrate that Semantic Features $\mathcal{F}_{\text{semantic}}$ and Acoustic Features $\mathcal{F}_{\text{acoustic}}$ are critical for speech processing tasks. This design choice helps focus on acoustic information, including $\mathcal{P}$ (pitch), $\mathcal{T}$ (timbre) and content information extracted from semantic tokens $\mathbf{S} = \{s_1, s_2, ..., s_n\}$.

For semantic information processing, we leverage the Whisper-large encoder [28] $\Phi_{\text{whisper}} : \mathcal{X} \rightarrow \mathcal{H}_s$, pre-trained through weak supervision, enabling robust extraction of linguistic features with zero-shot capabilities. The acoustic stream

utilizes WavLM [29] $\Phi_{\text{wavlm}} : \mathcal{X} \to \mathcal{H}_a$, which specializes in capturing speaker-specific characteristics and timbral qualities. This complementary combination allows our model to analyze content-level semantic features and fine-grained acoustic properties crucial for TB detection.

The processing pipeline operates on batched input signals $\mathbf{X} \in \mathbb{R}^{B\times T}$, where B represents the batch size and T the temporal dimension. For semantic analysis, the input is transformed into log-mel spectrograms $\mathbf{S} \in \mathbb{R}^{B\times F\times T}$, with F frequency bins. For acoustic analysis, we use raw audio waveforms to pass in the WavLM encoder. These representations are processed through their respective encoders to produce feature spaces:

$$\mathbf{H}_s = \Phi_{\text{whisper}}(\mathbf{S}) \in \mathbb{R}^{B\times T_s\times D_s}$$

$$\mathbf{H}_a = \Phi_{\text{wavlm}}(\mathbf{X}) \in \mathbb{R}^{B\times T_a\times D_a}$$

where T_s, T_a represent temporal dimensions and D_s, D_a denote feature dimensions for semantic and acoustic features respectively. This parallel processing enables our model to capture both semantic content and acoustic characteristics simultaneously. We pass the outputs $\mathbf{H}_s$ and $\mathbf{H}_a$ into semantic and acoustic adapters, respectively:

$$\mathbf{A}_s = f_s(\mathbf{H}_s) = \text{Adapter}_s(\mathbf{H}_s)$$

$$\mathbf{A}_a = f_a(\mathbf{H}_a) = \text{Adapter}_a(\mathbf{H}_a)$$

The semantic and acoustic adapters each comprise two sequential 1-D convolutional layers designed for temporal downsampling and alignment, followed by a down-up bottleneck adapter. Each adapter concludes with a linear projection layer $W_s \in \mathbb{R}^{D_s\times 1024}$ and $W_a \in \mathbb{R}^{D_a\times 1024}$ that maps the features to a shared dimensional space, enabling effective integration of both information streams. These adapters align both feature streams to a common representation space $\mathbb{R}^{B\times m\times 1024}$. We concatenate the outputs of both adapters to get:

$$\mathbf{x}_m = [\mathbf{A}_s; \mathbf{A}_a] \in \mathbb{R}^{B\times m\times 2048}$$

before passing them to the backbone for final classification.

3.2 Backend Network Architecture

The backend network of our TB detection system is built upon the RawNet3 architecture [25]. This architecture was selected for its proven effectiveness in capturing nuanced acoustic characteristics while maintaining computational efficiency crucial for medical applications with limited dataset availability.

At the core of our backend network are three AFMS-Res2MP blocks [25] arranged sequentially. Each AFMS-Res2MP block, as illustrated in Figure 2, consists of a sophisticated arrangement of convolutional layers with residual connections and feature map scaling. The input to each block first passes through a Conv1D layer, followed by ReLU activation and batch normalization (BN).

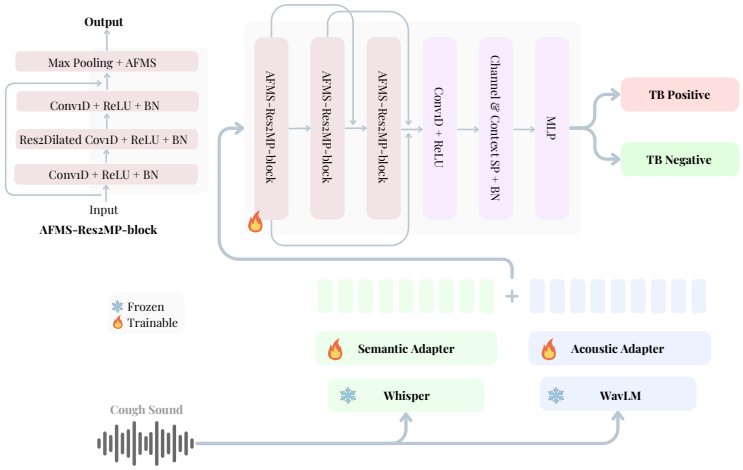

Fig. 2: Illustrates the architecture of the *AcouSem-AFNet* model. The proposed framework uses Whisper and WavLM encoders for semantic and acoustic processing, combined through adapters and AFMS blocks for final classification.

This is succeeded by a Res2Dilated Conv1D layer with ReLU and BN, which employs dilated convolutions to expand the receptive field without increasing the parameter count essential for capturing the broad temporal patterns in respiratory sounds indicative of TB. Another Conv1D layer with ReLU and BN follows, with its output added to the block input via a residual connection. The final component is a max pooling layer with an α Filter-wise feature Map Scaling (AFMS) module.

The AFMS module builds upon the filter-wise Feature Map Scaling (FMS) technique, which independently scales each filter of a feature map to derive more discriminative representations of respiratory characteristics. Let $c = [c_1, c_2, \cdots, c_F]$ represent a feature map output from a residual block, where $c_f \in \mathbb{R}^T$, T is the sequence length in time, and F denotes the number of filters. The AFMS module first performs global average pooling on the time axis, followed by feed-forwarding through a fully-connected layer with sigmoid activation to derive a scale vector $s = [s_1, s_2, \cdots, s_F]$, where $s_f \in \mathbb{R}^1$.

In our implementation for TB detection, we utilize both multiplicative and additive scaling methods to enhance the discriminative power of respiratory sound features. The multiplicative method applies the scale vector as $c'_f = c_f \cdot s_f$, while the additive method implements $c'_f = c_f + s_f$. For optimal performance in capturing the subtle acoustic signatures of TB, we employ a sequential application of both methods, expressed as:

$$c'_f = (c_f + s_f) \cdot s_f \tag{1}$$

This approach provides several advantages for TB detection from respiratory sounds. The multiplicative scaling functions similarly to an attention mechanism in the filter domain, allowing the network to emphasize specific frequency components particularly relevant to TB-related acoustic patterns. Unlike con-

ventional softmax-based attention, our sigmoid-based scaling prevents excessive information removal, preserving the complementary features captured by different filters indicating TB presence.

The additive scaling introduces data-driven perturbation to the feature maps, potentially increasing their discriminative power for distinguishing TB-related sounds from normal respiratory patterns or other conditions. This concept is particularly relevant for TB detection, where subtle variations in acoustic characteristics can significantly impact diagnostic accuracy. In the first two AFMS-Res2MP blocks, we incorporate max-pooling operations that serve as effective regularization techniques, crucial for preventing overfitting on our limited TB audio dataset. This design enhances the model's generalization capabilities across diverse patient populations and variable recording conditions while maintaining sensitivity to TB-specific respiratory signatures.

The three AFMS-Res2MP blocks process the concatenated features from the front-end dual streams, progressively refining the representation to capture the complex acoustic patterns associated with TB. A fully connected layer follows these blocks, mapping the extracted features to the final binary classification output indicating the presence or absence of TB. This carefully designed backend architecture enables robust TB detection from respiratory audio recordings while maintaining exceptional performance despite limited training data availability. The complete architecture of the *AcouSem-AFNet* model is shown in Figure 2.

4 Experimental Setup

Dataset: To assess the performance of the *AcouSem-AFNet* model for automated diagnosis of TB from cough sounds, we used the Cough Diagnostic Algorithm for Tuberculosis (CODA TB) challenge dataset] [30].The CODA-TB challenge, organized by Sage Bionetworks, focused on developing algorithms to detect pulmonary tuberculosis (TB) through acoustic analysis of cough recordings collected over a two-week period. This initiative aimed to advance automated TB screening by leveraging machine learning approaches to identify distinctive acoustic signatures in TB-positive cough episodes. The subject age is above 18 or older, with a cough sound of 0.5 seconds, collected from health centers of 7 countries (India, Philippines, South Africa, Uganda, Vietnam, Tanzania, Madagascar). The task associated is a TB-positive or TB-negative cough. The CODA-TB dataset contains 9,772 cough recordings from 1,105 patients, including detailed demographic, clinical, and microbiological diagnostic metadata. Initially developed for the CODA-TB DREAM Challenge competition, the training dataset is now publicly accessible[1]. Researchers can utilize this data to develop acoustic-based TB screening models, with model evaluation performed on a separate test set. To build our proposed approach, we performed the 70:10:20 train-val-test split on the publicly available training set.

[1] https://www.synapse.org/Synapse:syn31472953/wiki/619711

Baselines: We consider eight different existing architectures to compare the proposed *AcouSem-AFNet* model. We consider both variations of models that take raw audio as input and spectrogram-like features. We use models such as RawNet3 that take input as raw audio signals, whereas models such as MesoNet [31] and SpecRNet [32] take spectrogram features as input. We use the variant proposed in the paper [33] for the implementation of SpecRNet and MesoNet. For the spectrogram-based models, we use LFCC and the output of the Whisper ASR encoder.

Implementation Details: PyTorch is used to implement the proposed model. WavLM and Whisper are used from the Hugging Face repository. We use a variant of Whisper and WavLM. The proposed model is trained for 50 epochs, with AdamW as the optimizer and a batch size of 32. The learning rate and weight decay are set to 1e-3 and 5e-4, respectively. The training, validation, and evaluation data consist of the 6841, 977, and 1954 samples, respectively.

Evaluation Metrics: We evaluate our proposed *AcouSem-AFNet* approach using the AUROC and overall accuracy as the evaluation metrics.

5 Results

To assess the classification capabilities for TB diagnosis from cough acoustics, we compared the proposed *AcouSem-AFNet* model with the baseline models. The quantitative results are tabulated in Table 1, representing the AUC score and overall accuracy for baselines and the proposed model. From the results, the proposed approach outperforms the baseline models with an overall improvement from (3-18)% and (2-7)% in terms of the AUC and accuracy scores, respectively. To demonstrate performance, we provide the Receiver Operating Characteristics (ROC) Curve plot, shown in Figure 3.

Our evaluation demonstrates that the dual-stream architecture of *AcouSem-AFNet* outperforms all baseline models in tuberculosis detection from respiratory audio recordings. As shown in Table 1, the proposed model achieves superior performance with 78.10% accuracy and a mean AUC of 0.79, representing significant improvements over other baseline architectures.

The performance gap between our *AcouSem-AFNet* approach and the nearest competitor (SpecRNet with 76.61% accuracy and 0.76 AUC) highlights the effectiveness of our architectural innovations. While SpecRNet demonstrates strong performance as a single-stream approach, it fails to capture the full spectrum of TB-related acoustic signatures. Similarly, RawNet3 performs reasonably well (73.69% accuracy, 0.71 AUC) but lacks the specialized dual pathway design that enables comprehensive feature extraction for semantics and acoustic characteristics. The Whisper variants (SpecRNet and MesoNet) consistently underperform with accuracies around 71.6%, and AUC scores between 0.61-0.66, indicating their limitations in capturing the subtle acoustic patterns associated with TB respiratory sounds. This underscores the insufficiency of relying solely on general

Models	Accuracy(%)	Mean AUC
Whisper SpecRNet	71.55	0.66
SpecRNet	<u>76.61</u>	<u>0.76</u>
Whisper LCNN	71.60	0.65
MesoNet	74.41	0.74
Whisper MesoNet	71.65	0.61
RawNet3	73.69	0.71
AcouSem-AFNet	**78.10**	**0.79**

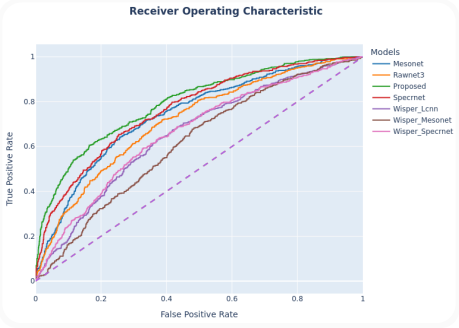

Fig. 3: TB detection model evaluation. **Left**: Performance metrics across architectures with AcouSem-AFNet (bold) achieving the best accuracy (78.10%), and AUC (0.79), second-best underlined. **Right**: ROC curves showing our model's (green) superior sensitivity-specificity balance versus baselines.

audio representation models for this specialized medical diagnostic task. The superior performance of *AcouSem-AFNet* model can be attributed to several key innovations: (1) The dual-stream architecture effectively leverages complementary information pathways, with the semantic stream (Whisper-large encoder) capturing structured respiratory patterns while the acoustic stream (WavLM) preserves critical temporal characteristics specific to TB-related sounds, (2) the specialized integration backbone with squeeze-excitation mechanisms and residual connections successfully fuses these complementary features while maintaining discriminative power, even with limited medical training data, and (3) the AFMS-Res2MP blocks with targeted regularization in the backend network effectively prevent overfitting while preserving the model's ability to detect subtle TB acoustic signatures.

Our AcouSem-AFNet model demonstrates balanced gender performance with an overall accuracy of 78.10%, achieving 76.01% for male subjects and 80.14% for female subjects. This relatively small performance gap suggests that the model effectively captures TB-related acoustic signatures which are independent of gender-based physiological differences in cough characteristics. Regarding age-related performance, we created four age bins to assess the generalizability of the model: 18-34, 35-51, 52-68, and 69-85. Performance analysis reveals interesting variations, with middle-aged groups (35-51 and 52-68) showing higher accuracy (80.4% and 80.7%, respectively) compared to younger (75.6%) and elderly populations (69.7%). However, we observed a concerning trend in precision and recall metrics for older age groups, particularly in the 52-68 bin, where precision drops to 40.6% and recall to 40%, indicating greater difficulty in correctly identifying TB-positive cases among older patients.

Clinical Significance: The 3% improvement in AUC and 2% gain in overall accuracy compared to the best baseline represent substantial progress in respiratory-based TB screening capabilities. These improvements could translate to meaningful clinical outcomes by reducing false negatives in TB screening protocols, potentially enabling earlier intervention and treatment.

Our experimental results conclusively demonstrate that the proposed *AcouSem-AFNet*, a dual-stream architecture with specialized feature integration, outperforms existing approaches for TB detection from respiratory audio recordings. The performance gains validate our hypothesis that capturing both semantic and acoustic features in cough sounds is essential for accurate TB diagnosis, opening promising avenues for non-invasive, cost-effective screening tools in resource-limited settings.

6 Conclusion

In this paper, we presented a novel deep-learning framework for automated TB detection through cough sound analysis. Our approach demonstrated superior performance on the CODA-TB challenge dataset, achieving state-of-the-art results in classification accuracy and robustness. The key innovations include our spectral representation technique and architecture optimization for acoustic biomarker detection. Experimental results validate our method's effectiveness as a rapid, non-invasive TB screening tool, particularly valuable for resource-limited settings. While our work shows promising results, future research directions include: (1) investigating multi-modal integration with clinical metadata, (2) improving model interpretability for clinical adoption, and (3) validating performance across diverse demographic populations. Our contribution establishes a strong foundation for acoustic-based TB screening, potentially addressing critical gaps in global TB detection and management.

Acknowledgment: This work is supported by Srijan: Center of Excellence on GenAI at IIT Jodhpur, India, IndiaAI Mission, and Meta.

Disclosure of Interests: The authors declare no competing interests.

References

1. WHO. [Accessed 23-02-2025].
2. I. Comas, M. Coscolla, *et al.*, "Out-of-Africa Migration and Neolithic Coexpansion of Mycobacterium Tuberculosis with Modern Humans," *Nature Genetics*, vol. 45, no. 10, pp. 1176–1182, 2013.
3. T. Paulson, "Epidemiology: A mortal foe.," *Nature*, vol. 502, no. 7470, 2013.
4. W. H. Organization, *Global tuberculosis report 2018*. World Health Organization, 2018.
5. G. Sulis and M. Pai, "Missing Tuberculosis Patients in the Private Sector: Business as usual will not deliver results," *Public Health Action*, vol. 7, no. 2, pp. 80–81, 2017.
6. R. R. Nathavitharana, C. Yoon, *et al.*, "Guidance for Studies Evaluating the Accuracy of Tuberculosis Triage Tests," *The Journal of Infectious Diseases*, vol. 220, pp. S116–S125, 2019.
7. Y. Akhter, R. Singh, and M. Vatsa, "AI-Based Radiodiagnosis using Chest X-rays: A Review," *Frontiers in Big Data*, vol. 6, p. 1120989, 2023.

8. P. Naidoo, G. Theron, *et al.*, "The South African Tuberculosis Care Cascade: Estimated Losses and Methodological Challenges," *The Journal of Infectious Diseases*, vol. 216, no. suppl_7, pp. S702–S713, 2017.
9. B. Simonsson, F. Jacobs, J. Nadel, *et al.*, "Role of Autonomic Nervous System and the Cough Reflex in the Increased Responsiveness of Airways in Patients with Obstructive Airway Disease," *The Journal of Clinical Investigation*, vol. 46, no. 11, pp. 1812–1818, 1967.
10. K. F. Chung and I. D. Pavord, "Prevalence, Pathogenesis, and Causes of Chronic Cough," *The Lancet*, vol. 371, no. 9621, pp. 1364–1374, 2008.
11. J. Korpáš, J. Sadloňová, and M. Vrabec, "Analysis of the Cough Sound: an Overview," *Pulmonary Pharmacology*, vol. 9, no. 5-6, pp. 261–268, 1996.
12. R. V. Sharan, U. R. Abeyratne, *et al.*, "Predicting Spirometry Readings using Cough Sound Features and Regression," *Physiological Measurement*, vol. 39, no. 9, p. 095001, 2018.
13. R. Pathri, S. Jha, *et al.*, "Acoustic Epidemiology of Pulmonary Tuberculosis (TB) & COVID-19 leveraging AI/ML," *medRxiv*, 2022.
14. M. Pahar, M. Klopper, *et al.*, "Automatic Cough Classification for Tuberculosis Screening in a Real-world Environment," *Physiological Measurement*, vol. 42, p. 105014, Nov 2021.
15. P. A. Marsden, I. Satia, *et al.*, "Objective Cough Frequency, Airway Inflammation, and Disease Control in Asthma," *Chest*, vol. 149, no. 6, pp. 1460–1466, 2016.
16. C. Infante, D. Chamberlain, *et al.*, "Use of Cough Sounds for Diagnosis and Screening of Pulmonary Disease," in *2017 IEEE Global Humanitarian Technology Conference (GHTC)*, pp. 1–10, 2017.
17. A. Windmon, M. Minakshi, *et al.*, "Tussiswatch: A Smart-Phone System to Identify Cough Episodes as Early Symptoms of Chronic Obstructive Pulmonary Disease and Congestive Heart Failure," *IEEE JBHI*, vol. 23, no. 4, pp. 1566–1573, 2018.
18. H. Sotoudeh, M. Tabatabaei, *et al.*, "Artificial Intelligence Empowers Radiologists to Differentiate Pneumonia Induced by COVID-19 versus Influenza Viruses," *Acta Informatica Medica*, vol. 28, no. 3, p. 190, 2020.
19. M. Pahar, M. Klopper, *et al.*, "Automatic Tuberculosis and COVID-19 Cough Classification using Deep Learning," in *2022 International Conference on Electrical, Computer and Energy Technologies*, pp. 1–9, 2022.
20. V. Ramesh, K. Vatanparvar, *et al.*, "CoughGAN: Generating Synthetic Coughs that improve Respiratory Disease Classification," in *IEEE EMBC*, pp. 5682–5688, 2020.
21. B. H. Tracey, G. Comina, *et al.*, "Cough Detection Algorithm for Monitoring Patient Recovery from Pulmonary Tuberculosis," in *IEEE EMBC*, pp. 6017–6020, 2011.
22. S. Larson, G. Comina, *et al.*, "Validation of an Automated Cough Detection Algorithm for Tracking Recovery of Pulmonary Tuberculosis Patients," *PLOS ONE*, vol. 7, pp. 1–10, 10 2012.
23. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *CVPR*, pp. 770–778, 2016.
24. G. Frost, G. Theron, and T. Niesler, "TB or not TB? Acoustic Cough Analysis for Tuberculosis Classification," *ArXiv:2209.00934*, 2022.
25. J. weon Jung, Y. Kim, *et al.*, "Pushing the Limits of Raw Waveform Speaker Recognition," in *Proc. Interspeech*, pp. 2228–2232, 2022.
26. S. Hu, L. Zhou, *et al.*, "WavLLM: Towards Robust and Adaptive Speech Large Language Model," in *Findings of the Association for Computational Linguistics: EMNLP*, pp. 4552–4572, 2024.

27. Z. Ye, P. Sun, *et al.*, "Codec Does Matter: Exploring the Semantic Shortcoming of Codec for Audio Language Model," *ArXiv:2408.17175*, 2024.
28. A. Radford, J. W. Kim, *et al.*, "Robust Speech Recognition via Large-Scale Weak Supervision," *CoRR*, vol. abs/2212.04356, 2022.
29. S. Chen, C. Wang, *et al.*, "Wavlm: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 1505–1518, 2021.
30. S. Huddart, V. Yadav, *et al.*, "A Dataset of Solicited Cough Sound for Tuberculosis Triage Testing," *Scientific Data*, vol. 11, no. 1, p. 1149, 2024.
31. D. Afchar, V. Nozick, *et al.*, "MesoNet: A Compact Facial Video Forgery Detection Network," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–7, 2018.
32. P. Kawa, M. Plata, and P. Syga, "SpecRNet: Towards Faster and More Accessible Audio Deepfake Detection," in *2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pp. 792–799, 2022.
33. P. Kawa, M. Plata, M. Czuba, *et al.*, "Improved DeepFake Detection Using Whisper Features," in *Proc. INTERSPEECH 2023*, pp. 4009–4013, 2023.