Rapid Tuberculosis Detection With Time-Frequency Representation of Smartphone-Recorded Cough Sounds and Depthwise Separable CNN

Roneel V. Sharan, *Senior Member, IEEE*, Junhua Li, *Senior Member, IEEE*, Minji Lee, *Member, IEEE*, and Yoshiharu Yamamoto, *Member, IEEE*

***Abstract*— Tuberculosis (TB) is a major public health concern, particularly in resource-constrained settings where access to rapid diagnostic tools remains limited. Cough is a characteristic symptom of TB, and its acoustic characteristics can provide valuable diagnostic information. In this work, we study an automated approach for TB detection based on cough sounds recorded using smartphone microphones. Our method utilizes mel-spectrogram, a time-frequency representation of cough sounds which captures rich perceptually relevant spectral-temporal information. The mel-spectrogram forms input to a lightweight convolutional neural network architecture based on MobileNet, which employs depthwise separable convolutions for computational efficiency, making it well-suited for deployment on mobile devices. The dataset used for this study comprises 9,993 cough sounds from 119 clinically diagnosed TB and non-TB subjects. To address the class imbalance in the dataset, we incorporate inverse class frequency weighting during training. We employ a subject-independent 10-fold cross-validation strategy to evaluate our model. Our approach achieves a precision of 0.888, a recall of 0.845, and an F1 score of 0.866 in detecting TB subjects, outperforming baseline methods based on handcrafted and deep learning features. These promising results highlight the potential of cough sound analysis using mobile devices as a rapid, accessible screening tool for TB.**

I. INTRODUCTION

Tuberculosis (TB) is a contagious and potentially life-threatening disease caused by the bacterium Mycobacterium tuberculosis [1]. It primarily affects the lungs, though it can also impact other parts of the body, such as the kidneys, spine, and brain [2]. TB spreads through the air when an infected person coughs, sneezes, or talks, making it highly transmissible, especially in crowded or poorly ventilated environments.

According to the World Health Organization (WHO) [3], TB remains a major global health problem. In 2023, an estimated 10.8 million people worldwide fell ill with TB, and approximately 1.25 million people died due to the disease. TB is one of the leading causes of death from a single infectious agent globally.

This work was supported by the University of Essex. *(Corresponding author: Roneel V. Sharan.)*

Roneel V. Sharan and Junhua Li are with the School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, United Kingdom (e-mail: roneel.sharan@essex.ac.uk; junhua.li@essex.ac.uk).

Minji Lee is with the Department of Biomedical Software Engineering, The Catholic University of Korea, Bucheon 14662, South Korea (e-mail: minjilee@catholic.ac.kr).

Yoshiharu Yamamoto is with the Educational Physiology Laboratory, Graduate School of Education, The University of Tokyo, Tokyo 113-0033, Japan (e-mail: yamamoto@p.u-tokyo.ac.jp).

Although TB is treatable with a prolonged course of antibiotics [4], early detection is important for successful treatment and prevention of transmission [5], [6]. However, diagnosing TB remains challenging. The gold-standard diagnostic test for pulmonary TB is bacterial culture from sputum samples, which is accurate but slow, often taking several weeks to yield results [7]. To address the need for quicker diagnosis, WHO recommends rapid molecular diagnostic tests as the initial test for individuals with TB symptoms [3]. These tests provide faster results.

Despite these advances, there are limitations to current rapid diagnostic tests [8], especially in resource-limited settings. Many low- and middle-income countries, where over 80% of TB cases and deaths occur [3], face barriers such as the high cost of equipment, lack of laboratory infrastructure, and shortage of trained healthcare workers [9], [10]. These constraints highlight the pressing need for accessible, affordable, and scalable alternative diagnostic tools.

A persistent cough is one of the most common symptoms of TB [11]. TB affects respiratory physiology by damaging lung tissue, reducing airflow, and causing inflammation, all of which can influence the acoustic properties of a person's cough [12]. While clinicians may use cough characteristics as part of clinical assessment, accurately interpreting cough sounds requires significant expertise and is subject to human variability [13]. Objective analysis of cough sounds, on the other hand, offers a promising approach for rapid and scalable TB screening, particularly when enhanced by machine learning techniques and implemented on smartphones [12].

This study proposes an automated method for detecting TB using smartphone-based cough sound analysis. Smartphone-recorded cough sounds are first converted into mel-spectrograms, a time-frequency representation that reflects the human auditory perception of sound. These mel-spectrograms are then used as input to a depthwise separable convolutional neural network (DSCNN) pretrained on general audio tasks. Given the relatively small number of samples in the dataset, transfer learning is employed, and the network is fine-tuned for binary classification: TB versus non-TB. The DSCNN is based on the MobileNetV1 architecture, making it suitable for deployment on smartphones. To address class imbalance, inverse class frequency weights are applied during training. The model is evaluated using subject-independent cross-validation on a dataset comprising cough sounds from individuals clinically diagnosed with TB or non-TB.

The remainder of this paper is structured as follows: Sec-

TABLE I
SUMMARY OF THE DATASET

Parameter	TB	Non-TB	All
Number of participants	84	35	119
Number of coughs	6,882	3,111	9,993
Age (years) range	19–65	24–66	19–66
Age (years) average	35.9±9.8	39.3±10.3	36.9±10.0
Gender (Male/Female)	62/22	21/13	83/35

tion II describes the methodology, including dataset details, mel-spectrogram formation, DSCNN architecture, baseline methods, and evaluation metrics. Section III presents the experimental setup and results. Section IV discusses the findings and concludes the paper.

II. MATERIALS AND METHODS

A. Dataset

In this work, we utilize data from a TB screening study [14]. Audio recordings of cough sounds were collected from participants with pulmonary TB and those with non-TB related coughs at the Centre for Respiratory Diseases Research (CRDR), Kenya Medical Research Institute (KEMRI), Nairobi, Kenya. Participants were adult outpatients recruited from National Treatment Program clinics and other chest clinics. Individuals in the TB group were clinically diagnosed with pulmonary TB. Non-TB participants presented with respiratory symptoms unrelated to TB and tested negative on molecular diagnostic assays.

Cough sounds were recorded in a quiet room using a smartphone (Google Pixel 2) at a 44.1 kHz sampling rate and 16-bit resolution. Each participant was recorded for up to two hours to capture passive (natural) coughs. The recording environment was selected to minimize background noise, and participants were advised to limit speech during the session.

All audio recordings were annotated by trained human annotators using Audacity software. Cough segments were extracted and filtered to remove any sounds contaminated by background noise (e.g., speech, fan, door noise, sneezing, or other respiratory sounds) or waveform distortions. Each validated cough sound was processed to a uniform duration of 1 second. Segments longer than 1 second were split into multiple clips, while shorter ones were zero-padded to reach the required length. Cough sounds shorter than 0.1 seconds were discarded. Refer to [14] for more details on data collection, extraction, and annotation.

From the total collection, only data from participants who consented to share their audio recordings were included in our study. The dataset used for analysis comprised 9,993 cough sounds from 119 subjects (84 TB and 35 non-TB). Table I summarizes the dataset characteristics, including the number of subjects, number of coughs, age range and average, and gender distribution. While age and gender information were complete for TB subjects, two non-TB subjects lacked age data and one lacked gender information.

B. Mel-Spectrogram Computation

As shown in Fig. 1(a), to extract features from cough audio recordings, each 1-second waveform was transformed into a mel-spectrogram, a time-frequency representation that maps acoustic power across a perceptually motivated frequency scale. This mel-spectrogram is input to the pretrained DSCNN, with a fixed input size of 96×64, where 96 represents the number of time frames and 64 represents the number of mel frequency bands.

Each audio signal, originally sampled at 44.1 kHz, was resampled to 16 kHz. The short-time Fourier transform (STFT) was computed using a Hann window of 25 ms and a hop length of 10 ms. The STFT segments the audio into overlapping windows and calculates the Fourier transform for each segment.

The magnitude of the STFT was then passed through a mel filter bank consisting of 64 triangular filters covering the range from 0 Hz to 8 kHz. This transforms the linear frequency axis to the mel scale, which better corresponds to how the human ear perceives frequency. The mel frequency for a given linear frequency (in Hz) was computed using O'Shaughnessy's formula [15].

The mel-spectrogram was computed by applying the filter bank to the magnitude of the STFT and taking the logarithm of the result.

C. Neural Network

In this work, we utilize YAMNet [16], a type of DSCNN that classifies audio clips into sound event categories. It is based on the MobileNet architecture [17], a lightweight DSCNN designed for efficient execution on mobile and embedded devices. YAMNet was trained on the AudioSet dataset [18], which consists of over two million labeled 10-second sound clips from YouTube videos, covering a wide range of 521 audio classes, including speech, music, environmental sounds, and machine noise. While YAMNet is not specifically trained for cough sound classification, cough is one of the 521 audio classes it has been trained on.

As shown in Fig. 1(b), the input to the network is a 96×64 mel-spectrogram, which is passed through layers of depthwise separable convolutions. The core of YAMNet is a MobileNetV1-like architecture using depthwise separable convolutions [19], which factorize a standard convolution into a depthwise convolution (spatial filtering per channel) followed by a pointwise convolution, drastically reducing the model size and computation while maintaining performance. In the provided modified architecture, the input layer accepts 96×64 single-channel mel-spectrograms. The convolutional layers progressively extract features, with grouped (depthwise) convolutions at key layers, and a global average pooling layer [20] aggregates features across time-frequency dimensions.

At the classification stage, the network is fine-tuned for binary classification by replacing the final dense layer with a fully connected layer of size two, and substituting the original classification layer with one that applies inverse class

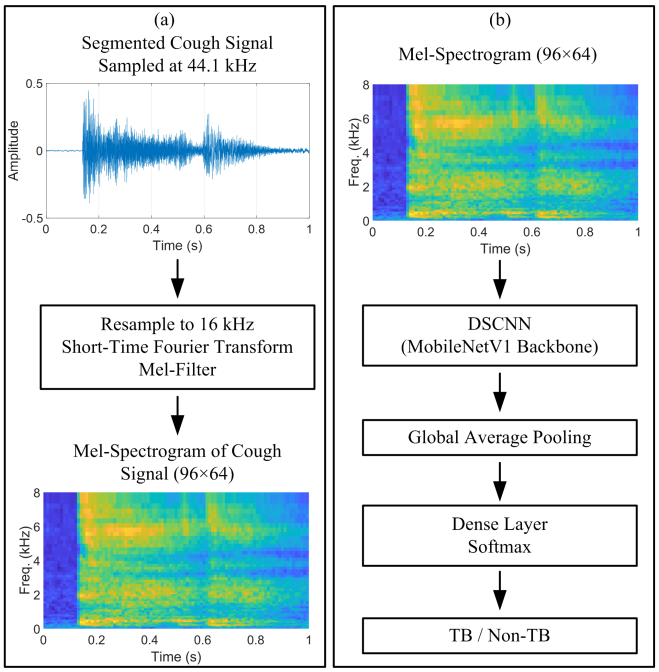

Fig. 1. Overview of (a) mel-spectrogram computation and (b) neural network architecture.

frequency weighting to address class imbalance. The class weights w_i are computed as

$$w_i = \frac{T}{KT_i}, \tag{1}$$

where T is the total number of training samples, K is the number of classes, and T_i is the number of samples for class i.

The final network has only about 3.2 M learnable parameters. The network is trained using the adaptive moment estimation optimizer [21] with a learning rate of 0.0003, for 20 epochs and a mini-batch size of 256.

D. Baseline Methods

To benchmark the proposed approach, we compare its performance against baseline methods that rely on handcrafted audio features and deep learning-based feature embeddings, combined with various machine learning classifiers [22].

The handcrafted feature sets are mel-frequency cepstral coefficients (MFCCs) and gammatone cepstral coefficients (GTCCs). MFCCs [23] capture the short-term power spectrum of audio using a mel-spaced filterbank, mimicking the human auditory system, while GTCCs are derived from gammatone filterbanks spaced on an equivalent rectangular bandwidth (ERB) scale [24], which better aligns with cochlear frequency selectivity. For each type, the first 13 static coefficients are extracted along with their first and second-order derivatives (delta and delta-delta) [25], resulting in a total of 39 features per type. These are computed over 25 ms Hamming windows with a hop length of 10 ms. The temporal statistics (mean and standard deviation) of each feature over the entire audio signal are computed, yielding a 78-dimensional feature vector for each cepstral feature set.

In addition, two pretrained audio embedding models are used to extract deep feature representations: VGGish and OpenL3. VGGish [16], based on a CNN architecture adapted from VGG, takes as input a 96×64 mel-spectrogram and outputs a 128-dimensional embedding vector representing high-level audio semantics. OpenL3 [26], a self-supervised embedding model trained for audio-visual correspondence, accepts a 128×199 mel-spectrogram input and produces a 512-dimensional embedding vector that captures detailed spectral-temporal characteristics of the audio signal.

The extracted features are evaluated using four commonly used supervised classification algorithms: logistic regression (LR), linear discriminant analysis (LDA), random forest (RF), and support vector machine (SVM) [27]. These methods represent a mix of linear and non-linear classifiers, enabling evaluation of both linearly and non-linearly separable patterns in the data.

To improve generalization and reduce feature dimensionality, two feature selection methods are considered. A two-sample t-test [28] is performed for each feature to assess whether its mean significantly differs between the two classes; features with p-values below 0.05 are retained. Also, an elastic net regularization approach [29] is used, which combines L_1 and L_2 penalties to perform both feature selection and shrinkage during model training, balancing sparsity and generalization.

E. Evaluation Metrics

To assess the performance of the proposed system and baseline methods, we use evaluation metrics commonly employed in binary classification tasks. These are precision, recall (or sensitivity), specificity, accuracy, F1 score, and area under the receiver operating characteristic (ROC) curve (AUC), with TB as the positive class and non-TB as the negative class.

III. EXPERIMENTAL EVALUATION

A. Experimental Setup

To evaluate the performance of the proposed and baseline methods, a 10-fold cross-validation approach was adopted. In each fold, data from 10% of the subjects were reserved for testing, ensuring subject independence. From the remaining 90% of the subjects, 90% were used for training and 10% for validation. This stratified setup maintains subject-level separation between training, validation, and test sets and provides a robust estimate of generalization performance. For the baseline methods, while we considered different feature selection methods, only the best results (highest F1 score and AUC) are presented here.

The proposed and baseline methods classify each cough as TB or non-TB. A subject is classified as having TB or non-TB using the cough index [30], which is the proportion of coughs classified as TB from each subject. Because of the limited number of subjects in each test fold, fold-wise computation of performance metrics can lead to high variance.

TABLE II
RESULTS USING BASELINE AND PROPOSED METHODS

Input	Classifier	Precision	Recall	Specificity	Accuracy	F1	AUC
MFCC	LR	0.795	0.690	0.571	0.655	0.739	0.703
MFCC	LDA	0.795	0.690	0.571	0.655	0.739	0.704
MFCC	RF	0.813	0.726	0.600	0.689	0.767	0.722
MFCC	SVM	0.842	0.762	0.657	0.731	0.800	0.754
GTCC	LR	0.827	0.738	0.629	0.706	0.780	0.725
GTCC	LDA	0.844	0.774	0.657	0.739	0.807	0.729
GTCC	RF	0.857	0.786	0.686	0.756	0.820	0.760
GTCC	SVM	0.811	0.714	0.600	0.681	0.759	0.741
VGGish	LR	0.808	0.702	0.600	0.672	0.752	0.692
VGGish	LDA	0.808	0.702	0.600	0.672	0.752	0.678
VGGish	RF	0.808	0.702	0.600	0.672	0.752	0.689
VGGish	SVM	0.808	0.702	0.600	0.672	0.752	0.715
OpenL3	LR	0.813	0.726	0.600	0.689	0.767	0.738
OpenL3	LDA	0.816	0.738	0.600	0.697	0.775	0.720
OpenL3	RF	0.792	0.679	0.571	0.647	0.731	0.701
OpenL3	SVM	0.842	0.762	0.657	0.731	0.800	0.766
Mel-Spectrogram	DSCNN	0.888	0.845	0.743	0.815	0.866	0.773

Therefore, predictions from all test folds are aggregated and performance metrics are computed at the subject level over the entire dataset, yielding more stable and representative results.

B. Experimental Results

Table II presents the TB vs non-TB subject classification results across multiple evaluation metrics using both baseline and proposed approaches. The results are reported for different feature extraction techniques and classifiers, including handcrafted cepstral features (MFCC and GTCC), deep learning features (VGGish and OpenL3), and the proposed DSCNN-based method using mel-spectrograms.

1) Results Using Cepstral Features: The classification performance using handcrafted cepstral features, MFCC and GTCC, shows notable differences in effectiveness depending on both the feature type and the choice of classifier. Across all classifiers, except SVM, GTCC features outperformed MFCCs, indicating that the ERB-scale filterbank used in GTCC provides a more perceptually relevant representation for this task. Among the classifiers, the SVM achieved the highest scores when using MFCCs, with a precision of 0.842, recall of 0.762, specificity of 0.657, accuracy of 0.731, F1 score of 0.800, and AUC of 0.754. RF followed closely, also delivering strong performance on MFCCs with an F1 score of 0.767.

However, GTCC features provided a performance boost across most classifiers. Using GTCC with RF yielded the best overall results in this category, with precision of 0.857, recall of 0.786, specificity of 0.686, accuracy of 0.756, F1 score of 0.820, and AUC of 0.760. LDA and LR also benefited from GTCCs, improving both precision and recall over their MFCC counterparts. These findings suggest that GTCCs are more effective for this audio classification task, and that nonlinear classifiers like RF capitalize better on the richer feature representation offered by GTCCs.

2) Results Using Deep Learning Features: The classification results using deep learning features, VGGish and OpenL3 embeddings, demonstrate the strength of learned audio representations across various classifiers. Between the two, OpenL3 features generally outperformed VGGish, indicating that the higher 512-dimensional OpenL3 embeddings offer a more discriminative feature space compared to the 128-dimensional VGGish features. For VGGish, all classifiers (LR, LDA, RF, and SVM) yielded identical precision (0.808), recall (0.702), specificity (0.600), accuracy (0.672), and F1 scores (0.752), with minor variations in AUC, suggesting that the feature representation was a limiting factor regardless of classifier complexity.

In contrast, OpenL3 provided more differentiation among classifiers. The SVM achieved the best performance with OpenL3, reaching a precision of 0.842, recall of 0.762, specificity of 0.657, accuracy of 0.731, F1 score of 0.800, and AUC of 0.766. LR and LDA also performed well with OpenL3, with LDA slightly ahead in terms of F1 score. Interestingly, RF showed a slight decline in performance with OpenL3, suggesting it may not fully exploit the structure in the high-dimensional feature space. Overall, OpenL3 features paired with SVM yielded the strongest deep learning feature-based results, highlighting the effectiveness of deep audio embeddings when used with appropriate classifiers.

3) Results Using Proposed Method: The proposed method, which utilizes mel-spectrogram inputs and a DSCNN architecture fine-tuned for binary classification, achieved the best overall performance across all metrics. It attained a precision of 0.888, recall of 0.845, specificity of 0.743, accuracy of 0.815, F1 score of 0.866, and an AUC of 0.773. These results highlight the effectiveness of end-to-end learning using task-specific model trained directly on time-frequency representations, outperforming handcrafted and deep learning features.

IV. DISCUSSION AND CONCLUSION

In this work, we investigated the detection of cough sounds associated with TB using a combination of mel-spectrogram and a lightweight DSCNN based on MobileNet architecture. We evaluated this approach against baseline methods using both handcrafted features (MFCCs and GTCCs) and deep learning audio embeddings (VGGish and OpenL3). Our experiments demonstrated that the proposed mel-spectrogram and DSCNN-based system outperformed the baseline methods, achieving the highest precision (0.888), recall (0.845), and F1 score (0.866). Among the baseline methods, GTCC features generally outperformed MFCC features, with RF and SVM classifiers providing the best performance across both handcrafted and deep learning feature sets. Deep learning embeddings, particularly OpenL3, offered improvements over handcrafted features, though they fell short of the proposed DSCNN-based approach.

The potential application of this work is particularly promising in the context of rapid, accessible TB screening. By leveraging cough sounds recorded via smartphones and processing them with a lightweight MobileNet-based network, this approach can enable non-invasive and real-time TB risk assessment in low-resource settings. Such a tool could be integrated into mobile health applications to assist healthcare professionals or even be used directly by individuals in the community, supporting early detection and treatment efforts for TB. Future work could explore deployment and real-world validation of the system, as well as expanding its capabilities to detect other respiratory diseases, ultimately contributing to global health initiatives for combating infectious diseases.

REFERENCES

[1] I. Smith, "Mycobacterium tuberculosis pathogenesis and molecular determinants of virulence," *Clinical Microbiology Reviews*, vol. 16, no. 3, pp. 463–496, 2003.

[2] N. Mohammadnabi, J. Shamseddin, M. Emadi, A. B. Bodaghi, M. Varseh *et al.*, "Mycobacterium tuberculosis: The mechanism of pathogenicity, immune responses, and diagnostic challenges," *Journal of Clinical Laboratory Analysis*, vol. 38, no. 23, 2024, Art. no. e25122.

[3] World Health Organization, "Tuberculosis," 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/tuberculosis

[4] V. A. Dartois and E. J. Rubin, "Anti-tuberculosis treatment strategies and drug development: Challenges and priorities," *Nature Reviews Microbiology*, vol. 20, pp. 685–701, 2022.

[5] J. Yayan, K.-J. Franke, M. Berger, W. Windisch, and K. Rasche, "Early detection of tuberculosis: A systematic review," *Pneumonia*, vol. 16, 2024, Art. no. 11.

[6] R. Bouatay, N. Bouaziz, H. Abdallah, S. Ben Hammouda, and J. Koubaa, "Isolated laryngeal tuberculosis: A diagnostic dilemma," *International Journal of Surgery Case Reports*, vol. 116, 2024, Art. no. 109376.

[7] T. A. Campelo, P. R. Cardoso de Sousa, L. d. L. Nogueira, C. C. Frota, and P. R. Zuquim Antas, "Revisiting the methods for detecting mycobacterium tuberculosis: What has the new millennium brought thus far?" *Access Microbiology*, vol. 3, no. 8, 2021.

[8] T. H. Divala, E. L. Corbett, C. Kandulu, B. Moyo, P. MacPherson *et al.*, "Trial-of-antibiotics to assist tuberculosis diagnosis in symptomatic adults in Malawi (ACT-TB study): A randomised controlled trial," *The Lancet Global Health*, vol. 11, no. 4, pp. e556–e565, 2023.

[9] L. M. Parsons, Ákos Somoskövi, C. Gutierrez, E. Lee, C. N. Paramasivan *et al.*, "Laboratory diagnosis of tuberculosis in resource-poor countries: Challenges and opportunities," *Clinical Microbiology Reviews*, vol. 24, no. 2, pp. 314–350, 2011.

[10] C. K. Shaozae, D. Das, and M. Kumar, "Diagnosis of tuberculosis in low-resource settings: Overcoming challenges within laboratory practice," *European Medical Journal*, 2023.

[11] G. O. Lee, G. Comina, G. Hernandez-Cordova, N. Naik, O. Gayoso *et al.*, "Cough dynamics in adults receiving tuberculosis treatment," *PLOS ONE*, vol. 15, no. 6, pp. 1–13, 2020.

[12] A. J. Zimmer, C. Ugarte-Gil, R. Pathri, P. Dewan, D. Jaganath *et al.*, "Making cough count in tuberculosis care," *Communications Medicine*, vol. 2, 2022, Art. no. 83.

[13] J. A. Smith, H. L. Ashurst, S. Jack, A. A. Woodcock, and J. E. Earis, "The description of cough sounds by healthcare professionals," *Cough*, vol. 2, 2006, Art. no. 1.

[14] M. Sharma, V. Nduba, L. N. Njagi, W. Murithi, Z. Mwongera *et al.*, "TBscreen: A passive cough classifier for tuberculosis screening with a controlled dataset," *Science Advances*, vol. 10, no. 1, 2024, Art. no. eadi0282.

[15] D. O'Shaughnessy, *Speech Communication: Human and Machine*, ser. Addison-Wesley series in electrical engineering. Addison-Wesley Publishing Company, 1987.

[16] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen *et al.*, "CNN architectures for large-scale audio classification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 131–135.

[17] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv:1704.04861*, 2017.

[18] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence *et al.*, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[19] L. Sifre, "Rigid-motion scattering for image classification," PhD Thesis, Ecole Polytechnique, 2014.

[20] M. Lin, Q. Chen, and S. Yan, "Network in network," in *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, Y. Bengio and Y. LeCun, Eds., 2014.

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, Y. Bengio and Y. LeCun, Eds., 2015.

[22] R. V. Sharan, K. Qian, and Y. Yamamoto, "Automated cough sound analysis for detecting childhood pneumonia," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 1, pp. 193–203, 2024.

[23] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[24] R. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory Physiology and Perception*, Y. Cazals, K. Horner, and L. Demany, Eds. Pergamon, 1992, pp. 429–446.

[25] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 1, pp. 52–59, 1986.

[26] A. L. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3852–3856.

[27] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer-Verlag, 2009.

[28] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[29] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, 2005.

[30] R. V. Sharan, S. Berkovsky, D. F. Navarro, H. Xiong, and A. Jaffe, "Detecting pertussis in the pediatric population using respiratory sound events and CNN," *Biomedical Signal Processing and Control*, vol. 68, 2021, Art. no. 102722.