# Automated Cough Sound Analysis for Detecting Childhood Pneumonia

Roneel V. Sharan, *Senior Member, IEEE*, Kun Qian, *Senior Member, IEEE*, and Yoshiharu Yamamoto, *Member, IEEE*

***Abstract*— Pneumonia is one of the leading causes of death in children. Prompt diagnosis and treatment can help prevent these deaths, particularly in resource poor regions where deaths due to pneumonia are highest. Clinical symptom-based screening of childhood pneumonia yields excessive false positives, highlighting the necessity for additional rapid diagnostic tests. Cough is a prevalent symptom of acute respiratory illnesses and the sound of a cough can indicate the underlying pathological changes resulting from respiratory infections. In this study, we propose a fully automated approach to evaluate cough sounds to distinguish pneumonia from other acute respiratory diseases in children. The proposed method involves cough sound denoising, cough sound segmentation, and cough sound classification. The denoising algorithm utilizes multi-conditional spectral mapping with a multilayer perceptron network while the segmentation algorithm detects cough sounds directly from the denoised audio waveform. From the segmented cough signal, we extract various handcrafted features and feature embeddings from a pretrained deep learning network. A multilayer perceptron is trained on the combined feature set for detecting pneumonia. The method we propose is evaluated using a dataset comprising cough sounds from 173 children diagnosed with either pneumonia or other acute respiratory diseases. On average, the denoising algorithm improved the signal-to-noise ratio by 44%. Furthermore, a sensitivity and specificity of 91% and 86%, respectively, is achieved in cough segmentation and 82% and 71%, respectively, in detecting childhood pneumonia using cough sounds alone. This demonstrates its potential as a rapid diagnostic tool, such as using smartphone technology.**

***Index Terms*— Cough sound, deep learning features, denoising, pneumonia, segmentation.**

## I. Introduction

PNEUMONIA accounted for 740,180 (14%) of all deaths in children under 5 years old in 2019, making it the leading infectious cause of death among children globally [1]. Pneumonia poses a significant health threat to children globally, with a disproportionate burden of morbidity and mortality observed in resource-deprived regions, particularly in southern Asia and sub-Saharan Africa [1]. Notably, indigenous children in developed nations like Australia and Canada also experience a disproportionate impact from pneumonia [2], [3].

Early detection and prompt treatment of pneumonia is essential in reducing childhood deaths [4]. The clinical algorithm developed by the World Health Organization employs symptoms like cough, breathing difficulty, fever, and chest pain to categorize pneumonia in regions with limited resources [5]. However, the algorithm has poor specificity [6], [7] because other acute respiratory diseases can also have similar symptoms. Inaccurate diagnosis of pneumonia can result in delayed or inappropriate treatment, contributing to the misuse of antibiotics and driving antibiotics resistance [8], [9]. Chest radiography can be used for differential diagnosis [10], [11] but it is not readily available in resource poor regions. The aforementioned situation underscores the necessity for the development of new rapid diagnostic tests specifically designed for pneumonia.

Cough is a common symptom of acute respiratory illnesses. It involves three distinct phases, inspiratory, compressive, and expiratory, and is an important defensive mechanism for maintaining lung health [12]. The sound produced during a cough is closely linked to its physiological process. Different respiratory conditions can impact different regions of the respiratory system. Consequently, these pathological changes can manifest in the sound of a cough [13], providing valuable clues about the underlying respiratory disease [14], [15], [16].

We posit that cough sound analysis can serve as a valuable screening method for childhood pneumonia. However, differentiating cough sounds by parents or caregivers may be impractical, while clinical assessment relies on the expertise and training of clinicians [17]. To address these limitations, our objective in this study is to develop a computational method for detecting childhood pneumonia through cough sound analysis. Subject to further external validation and clinical evaluation and, if widely distributed, such as through a smartphone application, such an objective assessment tool could serve as a valuable screening tool for parents and caregivers. Additionally, it has the potential to be particularly useful in developing countries and remote communities where

Research supported in part by the Google Research Scholar Program, the Oracle for Research Project Award, the Ministry of Science and Technology of the People’s Republic of China with the STI2030-Major Projects 2021ZD0201900, the National Natural Science Foundation of China (No. 62272044), the Teli Young Fellow Program from the Beijing Institute of Technology, JSPS KAKENHI (No. 20H00569), the JST Mirai Program (No. 21473074), and the JST MOONSHOT Program (No. JPMJMS229B).

Roneel V. Sharan is with the Educational Physiology Laboratory, Graduate School of Education, The University of Tokyo, Tokyo 113-0033, Japan, and also with the Australian Institute of Health Innovation, Macquarie University, Sydney, NSW 2109, Australia (e-mail: sharan@g.ecc.u-tokyo.ac.jp, roneel.sharan@mq.edu.au).

Kun Qian is with the School of Medical Technology, Beijing Institute of Technology, Beijing 10081, China (e-mail: qian@bit.edu.cn).

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).

access to healthcare facilities and clinicians is challenging.

### A. Related Works

Despite the plausibility of detecting childhood pneumonia using the sound of cough, our literature search found a limited number of previous studies [6], [18], [19], [20], [21], [22]. Two studies [6], [18] on detecting childhood pneumonia through cough sound analysis employ conventional feature extraction, including the utilization of manually crafted features, and classification techniques. Such conventional classification methods have been superseded by neural network-based classification methods, including deep learning techniques. This is true even on small datasets using techniques such as transfer learning [22] and shallow networks [23]. In addition, their test results are reported on only 25 pneumonia and non-pneumonia subjects, which makes it difficult to see the generalizability of their method, and the cough sounds are manually segmented, which prevents deployment of their method as a fully automated diagnostic aid.

Two studies [19], [20] report their results on much larger datasets with automatically segmented cough sounds. However, similar to the previous two studies [6], [18], their method employs conventional feature engineering and classification methods. Also, the cough sound features are augmented with clinical symptoms which makes it difficult to determine the contribution of the cough features to their classification algorithm.

The remaining two studies [21], [22] employ neural network-based techniques. However, in one study [21], the automatic cough segmentation algorithm is based on a simple energy threshold method, which may not work well in the presence of background noise and non-cough sounds, as present in their recordings. In addition, they did not report the performance of their cough segmentation algorithm and the pneumonia vs non-pneumonia cough classification results are reported on a small subset of their overall dataset, without a clear explanation on the method followed in the selection of this subset. In the remaining study [22], the cough sounds are manually segmented which once again prevents deployment of their method. Both the studies [21], [22] utilize the same dataset which contains background noise. These can degrade the quality of the cough sounds but neither of these studies perform noise filtering or report the performance of their denoising method.

### B. Automated Cough Sound Analysis

In this work, we propose a fully automated method to detect childhood pneumonia that utilizes only cough sound characteristics. In particular, our method employs different neural network methods for denoising cough sounds, cough sound segmentation, and cough sound classification. In denoising cough sounds, compared to an earlier study [24], our work proposes multi-conditional training of a multilayer perceptron (MLP) model to emulate the different noise levels that can be present in any given environment and we evaluate the performance of our denoising algorithm on noisy cough sound recordings collected in a real-life environment.

Our cough segmentation algorithm learns directly from the denoised cough waveforms, without the need for manual feature engineering as seen in earlier works [19], [20], [21]. This is achieved using a convolutional neural network, with a customized first layer, that learns the spectral characteristics in small time windows and a recurrent neural network that learns the temporal dependencies between the windows. In distinguishing between pneumonia and non-pneumonia cough sounds, when compared to previous studies [6], [18], [19], [20], we incorporate a set of deep learning features extracted from a pretrained audio classification network to leverage the power of learned representations in analyzing cough sounds. In addition, we employ a MLP that is trained on a combined handcrafted and deep learning features to grasp intricate relationships between input features and amalgamate them into higher-level representations. We validate our method on a clinically annotated dataset that encompasses nearly twice the number of subjects compared to two previous studies [6], [18]. To our knowledge, this is the first comprehensive work that utilizes only cough sounds for detecting childhood pneumonia where the cough sound analysis is fully automated.

## II. Materials and Methods

An overview of the proposed approach is shown in Fig. 1 which comprises three main components: cough sound denoising, cough sound segmentation, and cough sound classification (pneumonia vs non-pneumonia). Background noises present in everyday environments can degrade the quality of the cough sounds. In this work, we draw inspiration from advancements in speech enhancement research [25], [26], to denoise cough sound recordings. Our approach utilizes supervised learning, employing a MLP, to establish a mapping between the spectra of noisy and clean cough sound signals.

The denoised recordings form input to a cough sound segmentation network to detect the start and end point of the cough sounds in each recording. For cough segmentation, we propose a method inspired by SincNet [27], which operates on the raw waveform of the denoised recordings. SincNet is a one-dimensional convolutional neural network that uses sinc functions in the first convolutional layer to discover more meaningful filters. In this work, we combine the SincNet with a bidirectional gated recurrent unit (BiGRU), a type of recurrent neural network [28]. In the combined SincNet-BiGRU network [29], the SincNet learns the spectral characteristics within each windowed signal and BiGRU learns the bidirectional temporal dependencies between the windows.

We extract features from the segmented cough sounds for distinguishing between pneumonia and non-pneumonia. In line with previous studies [6], [18], [19], [20], our work incorporates a variety of manually crafted features that capture different aspects of cough sounds. However, our work also utilizes a set of deep learning features and a MLP is employed for classification on the combined feature set. We assess the effectiveness of our proposed approach using a clinically verified dataset of cough sound recordings collected from children who have been diagnosed with pneumonia or other acute respiratory illnesses. More details on the components of Fig. 1 are discussed in the following subsections.

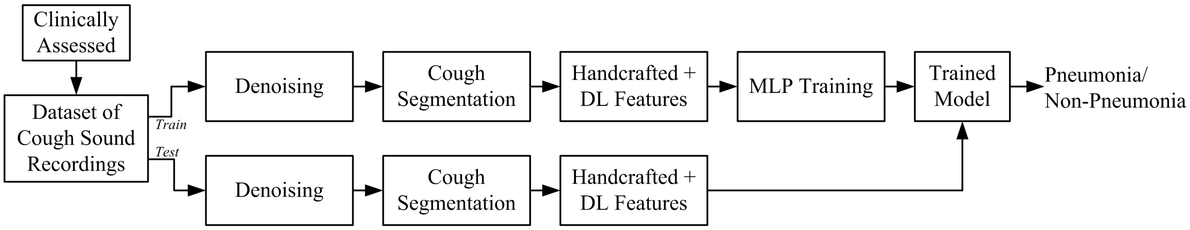

Fig. 1. An overview of the proposed method for distinguishing between pneumonia and non-pneumonia in children based on cough sound analysis.

TABLE I
OVERVIEW OF THE PNEUMONIA VS NON-PNEUMONIA DATASET USED IN THIS WORK

| | **Disease Group** | | |
|---|---|---|---|
| | **Pneumonia** | **Non-Pneumonia** | **Overall** |
| Number of subjects | 82 | 91 | 173 |
| Total duration (s) | 372.10 | 320.61 | 692.71 |
| Number of coughs | 268 | 223 | 491 |
| Gender (male:female) | 43:39 | 51:40 | 94:79 |
| Age range (years) | 0–11 | | |

TABLE II
OVERVIEW OF THE TRAINING DATASET FOR THE COUGH DENOISING ALGORITHM

| **Description** | **Value** |
|---|---|
| Number of recordings | 300 |
| Average duration (seconds) | 9.35±1.36 |
| Number of frames | 700,594 |
| Gender (male:female:unknown) | 169:90:41 |
| Average age (years) | 39.24±14.62 |

### *A. Dataset*

In this study, we utilized an open-source dataset of cough sound recordings obtained from West China Second University Hospital of Sichuan University [21]. Table I provides an overview of the dataset [22]. The dataset encompasses audio recordings of cough sounds from 173 children diagnosed with acute respiratory diseases, which can be classified into two categories: *pneumonia* and *non-pneumonia*. The pneumonia class consists of 82 subjects (43 male and 39 female), with 55 subjects diagnosed with pneumonia, 23 subjects with bronchopneumonia, and 4 subjects with lobar pneumonia. The non-pneumonia class comprises 91 subjects (51 male and 40 female), of which 80 subjects have acute bronchitis, 6 subjects have acute bronchiolitis, and 5 subjects have acute asthmatic bronchitis. The diagnosis of the diseases followed clinical guidelines [30]. The age range of the children included in the dataset is from 0 to 11 years, with the majority being one year or younger.

The cough sound recordings in this dataset are available in the MP3 file format, with a sampling frequency of 44.1 kHz. MP3 is a lossy conversion format where a psychoacoustic model utilizes the critical bands of human hearing to discard inaudible information from the audio [31]. However, the use of human auditory filters has shown to be useful in the analysis of cough sounds [14], [15]. For this reason, we downsampled all the recordings to 16 kHz for further processing. The recordings are captured within a hospital setting, incorporating background noises such as speech and sounds produced by medical devices. For the pneumonia class, the cumulative duration of the recordings amounts to 372.10 seconds, while for the non-pneumonia class, it totals 320.61 seconds. Before proceeding with our analysis, all the recordings underwent manual screening to ensure their suitability for this study. As a result, two pneumonia recordings are excluded due to the absence of any cough sounds. Additionally, one pneumonia recording is omitted since it was unclear whether the respiratory sounds are cough or non-cough. Each of the remaining recordings contains one or more cough sounds. The pneumonia class comprises a total of 268 cough sounds, while the non-pneumonia class comprises 223 cough sounds. All recordings are converted to the WAV file format for further processing. The waveforms of pneumonia and non-pneumonia (bronchitis) coughs are illustrated in Fig. 2(a) and (b), respectively.

The development of the deep learning-based denoising method employed in this work requires a mapping between a set of clean and noisy cough sound recordings. Since the dataset of pneumonia and non-pneumonia recordings used in this work is already noisy, we use 300 cough sound recordings, resampled at 16 kHz, from the COUGHVID dataset [32] for this purpose. In an earlier study [24], these 300 recordings were manually screened to ensure minimal to no background noise was present. A description of this dataset is provided in Table II. The recordings have an average duration of 9.35±1.36 seconds and have a total of 700,594 frames (frame length of 256 points with a 75% overlap between adjacent frames).

For the deep learning based cough segmentation algorithm, we employ a dataset for pretraining the network. The pretraining dataset has 300 cough and non-cough sound recordings taken from the COUGHVID dataset [29]. This dataset has been manually annotated to determine the location (start and end points) of the cough sounds in each recording. A description of this dataset is provided in Table III. The dataset has 150 cough and 150 non-cough recordings sampled at 16 kHz. These recordings are divided into frames of length 64 ms with 25% overlap between frames, resulting in a total of 55,332 frames for training the cough segmentation algorithm.

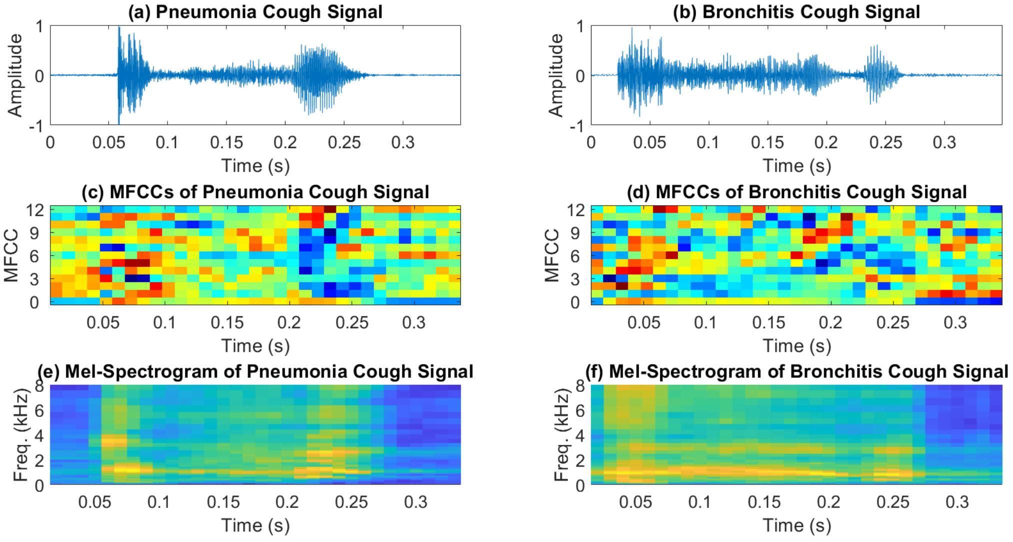

Fig. 2. The cough sound waveform for (a) pneumonia and (b) bronchitis, their mel-frequency cepstral coefficients in (c) and (d) respectively, and their mel-spectrogram representation in (e) and (f) respectively.

TABLE III
OVERVIEW OF THE COUGH AND NON-COUGH DATASET USED FOR PRETRAINING THE COUGH SEGMENTATION ALGORITHM

| | **Cough** | **Non-Cough** | **All** |
|---|---|---|---|
| Number of recordings | 150 | 150 | 300 |
| Average duration (s) | 8.76 | 9.04 | 8.90 |
| Number of frames | 27,229 | 28,103 | 55,332 |
| No. of coughs in recordings | 683 | 0 | 683 |
| Average age (years) | 35.57 | 38.00 | 36.09 |
| Gender (male:female:unknown) | 92:40:18 | 22:22:106 | 114:62:124 |

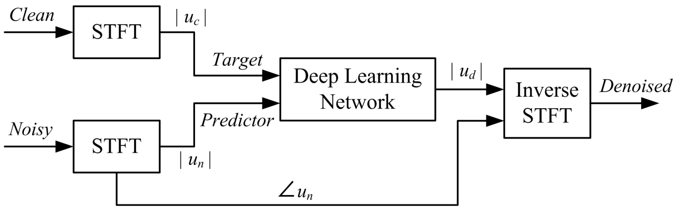

Fig. 3. Overview of the training procedure of the mapping-based cough denoising method.

### B. Cough Denoising

An overview of the mapping-based cough denoising method used in this work is outlined in Fig. 3. The main background noise type present in our evaluation dataset (Table I) is speech. As such, to create the noisy cough recordings, speech babble noise from the NOISEX-92 database [33] is added to the 300 clean cough recordings (Table II). These are added at a signal-to-noise ratio (SNR) of 20 dB, 10 dB, and 0 dB for multi-conditional training as the noise level varies in the evaluation dataset.

The predictor input to the MLP is formed by the magnitude spectra of the noisy cough recordings, while the target input is formed by the magnitude spectra of the clean cough recordings. Conversion to frequency domain is performed using short-time Fourier transform (STFT), applied using a window length of 256 points, 75% overlap between adjacent frames, and a Hamming window [24]. The resulting spectral vector is reduced to 129 by discarding the symmetric half. The MLP serves as a regression network, aiming to minimize the mean square error between its output and the target input, producing the magnitude spectrum of the denoised signal. The denoised cough signal is obtained by converting the denoised spectra back to the time domain using the phase of the noisy signal and the inverse STFT [25].

The MLP is a type of fully connected feedforward artificial neural network. The MLP architecture [24], Fig. 4, consists of an input layer, two hidden layers, and an output layer. The predictor input has a size of 129×8, as each prediction of the STFT output (129×1) is based on the current noisy STFT vector and the previous 7 STFT vectors. The predictor matrices and target vectors are normalized using their mean and standard deviation values. The two fully connected layers have 1024 neurons each. Each fully connected layer is succeeded by a batch normalization layer [34] and a rectified linear unit (ReLU) layer [35]. The output layers consist of a fully connected layer with a size of 129 (matching the target vector) and a regression layer. The network is trained using the adaptive moment estimation algorithm [36] with an initial learning rate of $1 \times 10^{-3}$, a mini-batch size of 128, and a maximum number of epochs of 3. Additionally, a learning rate drop factor of 0.9 and a learning rate drop period of 1 are employed.

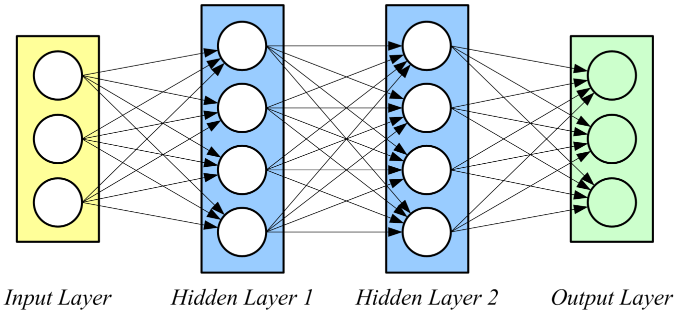

Fig. 4. An overview of the MLP architecture. The MLP has an input layer, two fully connected layers, and an output layer.

### C. Cough Segmentation

The denoising algorithm removes background noises from the recordings, however, the recordings can contain both cough and non-cough sound events, both speech and non-speech. This brings about the need for cough sound segmentation, that is, detecting the start and end points of the cough sounds in the denoised recordings, before performing cough sound analysis. The cough segmentation network [29], depicted in Fig. 5 with network architecture details in Table IV, operates on raw audio waveforms, with each input sequence consisting of 64 ms of audio data and a 25% overlap between sequences. Given a sampling frequency of 16 kHz, each sequence contains 1024 data points. The SincNet model applies convolutional operations independently to each time step of the audio signal sequences. The SincNet architecture consists of three sets of convolutional layers. The first layer employs sinc-based convolutions with 80 filters of length 251. These sinc functions implement bandpass filters with adjustable cutoff frequencies, which are learned during training. The convolution operation is carried out using a predefined function with a rectangular frequency response [27]. The initial values of the cutoff frequencies are set based on the equivalent rectangular bandwidth [37], which is a psychoacoustic measure representing the width of human auditory filters. The Sinc layer aims to optimize the parameters of these bandpass filters within the neural network framework. Consequently, this approach facilitates faster convergence during training and yields improved performance compared to standard CNNs [27].

The subsequent two convolutional layers are standard convolutions, utilizing 60 filters of length 5. Following each convolutional layer, there is a batch normalization layer, a leaky ReLU layer [38] with a negative input multiplier of 0.2, and a 1×3 max pooling layer. The stride for all convolutional and max pooling layers is set to 1. Following the convolutional layers, there are three fully connected layers, each with an output size of 256. Batch normalization and leaky ReLU layers follow each fully connected layer. A flatten layer is then employed to reshape the output into vector sequences. To capture bidirectional long-term dependencies between the time steps of the sequence data, a BiGRU [29] is utilized. This BiGRU layer learns the relationships in both the forward and backward directions. The final layers of the network consist of a fully connected layer, a softmax layer [39], and a classification layer.

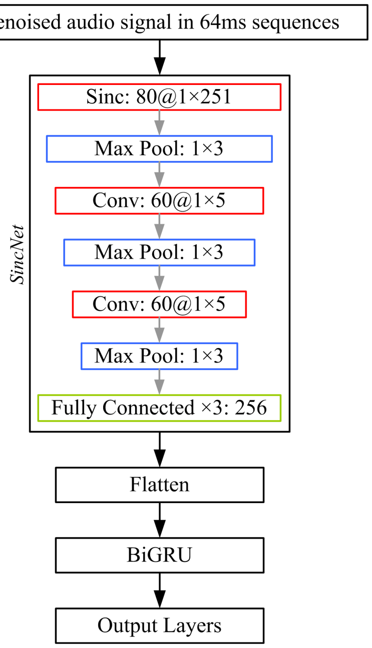

Fig. 5. The SincNet-BiGRU architecture for cough segmentation.

### D. Pneumonia vs Non-Pneumonia Cough Classification

As depicted in Fig. 1, the automatically segmented cough signals are processed to extract two sets of features: *hand-crafted features* and *deep learning features*.

*Handcrafted Features*: This study employs two types of handcrafted features, namely *cepstral* and *temporal-spectral* features. To compute these features, each segmented cough signal is divided into frames of 25 milliseconds, with a 15-millisecond overlap between adjacent frames. The cepstral features used are *mel-frequency cepstral coefficients (MFCCs)* [40], which are commonly employed in audio classification tasks. MFCCs utilize frequency scales based on auditory perception. In each frame, we calculate 13 MFCCs. A plot of the MFCCs, as a heatmap, for waveforms of pneumonia and non-pneumonia (bronchitis) coughs are shown in Fig. 2(c) and (d), respectively. In addition, we compute the first and second derivatives of the MFCCs [41]. This process results in a 39-dimensional feature vector for each cough frame. To represent these raw features, we employ the *mean* and *standard deviation* statistical measures. If a recording contains only one cough, these statistics are computed across all frames within that cough. If a recording comprises multiple coughs, these statistics are computed across all frames from all the coughs. As a result, we obtain a 78-dimensional MFCC feature subset for each recording.

The second subset of handcrafted features capture 15 temporal and spectral characteristics of the cough signal, as described in [22]. These features are the *zero-crossing rate*, *short-time energy*, *pitch*, *harmonic ratio*, and 11 *spectral* characteristics, which are the *spectral crest*, *centroid*, *entropy*, *decrease*, *flux*, *flatness*, *kurtosis*, *skewness*, *roll-off point*, *spread*, and *slope*. Similar to MFCCs, these features are computed in each frame and represented using the mean and standard deviation statistical measures. Consequently, each recording is

TABLE IV
DETAILS OF THE SINCNET-BIGRU ARCHITECTURE

| Layer Type | No. of Filters | Kernel | Stride | Output |
|---|---|---|---|---|
| Input | | | | 1×1024×1 |
| Sinc | 80 | 1×251 | | 1×774×80 |
| Batch Norm | | | | 1×774×80 |
| Leaky ReLU | | | | 1×774×80 |
| Max Pool | | 1×3 | 1 | 1×772×80 |
| Convolution | 60 | 1×5 | 1 | 1×768×60 |
| Batch Norm | | | | 1×768×60 |
| Leaky ReLU | | | | 1×768×60 |
| Max Pool | | 1×3 | 1 | 1×766×60 |
| Convolution | 60 | 1×5 | 1 | 1×762×60 |
| Batch Norm | | | | 1×762×60 |
| Leaky ReLU | | | | 1×762×60 |
| Max Pool | | 1×3 | 1 | 1×760×60 |
| Fully Connected | | | | 1×1×256 |
| Batch Norm | | | | 1×1×256 |
| Leaky ReLU | | | | 1×1×256 |
| Fully Connected | | | | 1×1×256 |
| Batch Norm | | | | 1×1×256 |
| Leaky ReLU | | | | 1×1×256 |
| Fully Connected | | | | 1×1×256 |
| Batch Norm | | | | 1×1×256 |
| Leaky ReLU | | | | 1×1×256 |
| Flatten | | | | 256 |
| BiGRU | | | | 512 |
| Fully Connected | | | | 2 |
| Softmax | | | | 2 |
| Output | | | | 2 |

associated with a 30-dimensional temporal and spectral feature subset.

*Deep Learning Features*: The deep learning feature set comprises 128 VGGish feature embeddings obtained from each cough signal. These embeddings are extracted using a pretrained convolutional neural network designed for audio classification [42]. The VGGish architecture draws inspiration from the popular VGG networks used in image classification tasks. It has been trained on a large dataset of YouTube audio, generating 128-dimensional embeddings. To compute the VGGish features, each cough signal is transformed into a mel-spectrogram, as depicted in Fig. 2. While Fig. 2 displays the mel-spectrogram of segmented cough signals, for input to the VGGish network, the signals are either zero-padded or cropped to a duration of 0.975 seconds before computing a 96×64 mel-spectrogram. The resulting mel-spectrogram, which represents the time-frequency characteristics of the cough signal, serves as the input to the VGGish network for extracting the feature embeddings. In the case of recordings containing multiple coughs, the feature embeddings are averaged across all the coughs.

The combined feature vector consists of 236 dimensions, encompassing 78 MFCC features, 30 temporal and spectral features, and 128 VGGish features. These features extracted from cough signals are employed for binary classification, distinguishing between pneumonia and non-pneumonia cases. The classification task utilizes three different classifiers: *random forest (RF)*, *support vector machine (SVM)*, and *multilayer perceptron (MLP)*. In classification, we consider all the features as input to the classifier but we also consider the most discriminative features, identified through the application of *t*-test and elastic net methods [43]. The MLP architecture is similar to Fig. 4, but now performing classification instead of regression. It consists of two hidden layers, each comprising 256 neurons and employing the rectified linear unit activation function. The network is trained using adaptive moment estimation with a learning rate of $3 \times 10^{-3}$, a mini-batch size of 8, and a maximum number of epochs of 10.

### E. Evaluation Metrics

The classification performance of the cough segmentation (cough vs non-cough) method and pneumonia vs non-pneumonia cough classification method is evaluated using sensitivity, specificity, accuracy, and $F_1$ score. These metrics are computed as

$$Sensitivity = \frac{TP}{TP + FN} \tag{1}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$F_1 = \frac{2TP}{2TP + FP + FN} \tag{4}$$

where *TP*, *TN*, *FP*, and *FN* represent the number of true positives, true negatives, false positives, and false negatives, respectively. We also use the area under the curve (AUC) of the receiver operating characteristic (ROC) curve as a single measure of classification performance.

For pneumonia vs non-pneumonia cough sound classification, the positive and negative classes are pneumonia and non-pneumonia, respectively. In the context of cough segmentation, the positive and negative classes correspond to cough and non-cough, respectively, and the metrics are computed similar to [44]. The reference label for cough and non-cough is based on manual segmentation, that is, we manually determined the start and end points of the cough sounds in each recording by visual analysis of its temporal waveform and the spectrogram representation using Audacity (www.audacityteam.org), a digital audio editing software. We perform frame-based classification in cough segmentation [29] whereby each recording is divided into 64 ms frames, with 25% overlap between adjacent frames, and a frame is labeled as cough if 50% or more data points in the frame contain cough and non-cough otherwise. When training the cough segmentation algorithm, these frame labels are used as the target or reference labels. When testing the cough segmentation algorithm, these frame labels are used to compute the evaluation metrics against the predicted labels.

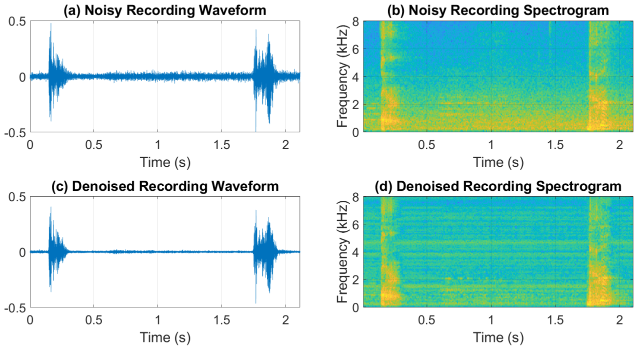

Fig. 6. Illustration of (a) waveform of recording with background noise, (b) spectrogram of recording with background noise, (c) waveform of denoised recording, and (d) spectrogram of denoised recording.

## III. Experimental Evaluation

### A. Cough Denoising

The cough denoising algorithm is trained and validated on the 300 recordings from the COUGHVID dataset using a similar procedure described in an earlier study [24], however, with multi-conditional training. We then evaluated the trained network on the pneumonia and non-pneumonia cough sound recordings. On average, the denoising algorithm improved the SNR by 43.56%. The improvement in the SNR for different noise levels is as follows: 117.38% improvement in SNR for recordings with SNR in the range of 0 – 10 dB, 45.71% improvement in SNR for recordings with SNR in the range of 10 – 20 dB, and 14.96% improvement in SNR for recordings with SNR in the range of 20 – 30 dB. As such, the most improvement in SNR is observed in recordings with high noise levels.

In Fig. 6, we provide illustration of a noisy and denoised cough recording waveform and their spectrogram representation. In this instance, the SNR improved from an estimated 15.21 dB to 23.54 dB. Visual analysis of the illustrations shows that the denoising algorithm largely maintains the temporal and spectral characteristics of the cough sounds while reducing background noise.

### B. Cough Segmentation

The SincNet-BiGRU classifies each 64 ms frame as a cough or non-cough frame. Connected frames with the same label are then categorized as cough or non-cough sound events. We first use the pretrained SincNet-BiGRU network to classify the cough and non-cough sequences, without retraining the network. Next, the performance of the SincNet-BiGRU network is evaluated in five-fold cross-validation, whereby, in each fold, frames from 80% of the pneumonia and non-pneumonia recordings are used for fine-tuning the network parameters and the frames from the remaining 20% of the recordings are used for testing. For training and validating the network using supervised learning and for computing the performance metrics, we use manual annotation of the cough sounds [22].

In Table V, we present the segmentation results using both, the pretrained network and the fine-tuned network. An accuracy of 0.7452, $F_1$ of 0.5662, and AUC of 0.8442 are achieved using the pretrained network, without any fine-tuning. This is in contrast to an evaluation accuracy of 0.9496, $F_1$ of 0.8476, and AUC of 0.9866 that could be achieved using this network on the COUGHVID dataset [29]. We believe there are two main reasons which could explain this discrepancy. Firstly, the COUGHVID dataset is crowdsourced with most recordings believed to be done in a home environment. While the dataset contains some background noise, the noise level is not as high as the dataset of pneumonia and non-pneumonia cough sounds used in this work which has been recorded in a noisy hospital environment. As such, both the noise environment and the noise levels are different. Secondly, the COUGHVID dataset comes from the adult population while the dataset used in this work comes from the pediatric population. The sound of cough is associated with the respiratory physiology but there are distinct differences in respiratory physiology between children and adults [45].

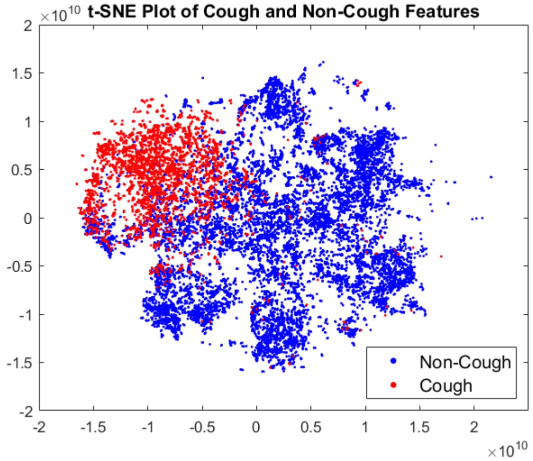

Fig. 7. *t*-SNE visualization of the SincNet-BiGRU network activations for cough vs non-cough classification.

Using fine-tuning, we are able to improve the accuracy in cough and non-cough frame classification to 0.8730 with $F_1$ of 0.7555 and AUC of 0.9498. This is an improvement of 17.15% in accuracy, 33.43% in $F_1$, and 12.51% in AUC over the results using the pretrained network (without fine-tuning). As such, the classification results are significantly improved after fine-tuning as the network acclimatizes to a different recording environment and a different population group.

Next, we examine the predictions of the SincNet-BiGRU model for cough and non-cough frames in the test set by employing *t*-distributed stochastic neighbor embedding (*t*-SNE) [46]. *t*-SNE is a technique that maps high-dimensional data, such as network activations, into two dimensions. The *t*-SNE visualization, depicted in Fig. 7, demonstrates that *t*-SNE is largely preserving the local structure of the data, that is, data points with similar characteristics are grouped together in the visualization. This implies that clusters or groups of data points share common classes.

TABLE V
COUGH SOUND SEGMENTATION RESULTS

| Input | SincNet-BiGRU Fine-Tuning | Classification Results | | | | |
|---|---|---|---|---|---|---|
| | | Sensitivity | Specificity | Accuracy | $F_1$ | AUC |
| Denoised signal | No | 0.7735 | 0.7375 | 0.7452 | 0.5662 | 0.8442 |
| Denoised signal | Yes | **0.9125** | **0.8622** | **0.8730** | **0.7555** | **0.9498** |

### C. Pneumonia vs Non-Pneumonia Cough Classification

The evaluation of the proposed method for classifying pneumonia vs non-pneumonia cough sounds involves a leave-one-out cross-validation approach. In this approach, the features from one subject are set aside for testing, while the features and class labels from the remaining subjects are used for training. This process is repeated for each subject to ensure every individual's data is used for testing once. Within each iteration, the features are standardized using $z$-score normalization. We report the results using the feature selection technique that yielded the best overall performance. In each fold, the discriminative features are determined based on the training data. For the $t$-test and elastic net feature selection methods, the discriminative features are chosen using a $p$-value threshold of 0.05 and the minimum cross-validated mean square error, respectively. The results for all classifiers are presented using three feature sets: handcrafted features, deep learning features, and a combined feature set.

We use leave-one-out cross-validation for the pneumonia vs non-pneumonia classifiers (RF, SVM, and MLP) because these classifiers are trained from scratch and using leave-one-out cross-validation allows us to maximize the training data for these classifiers. On the other hand, the cough segmentation network (SincNet-BiGRU) was pretrained on the COUGHVID dataset (Table III) and the network weights only required fine-tuning on the pneumonia vs non-pneumonia dataset. Also, training or fine-tuning the SincNet-BiGRU is much more time consuming then training the RF, SVM, and MLP classifiers. For these reasons we used 5-fold cross-validation for the cough segmentation algorithm.

The results for pneumonia vs non-pneumonia cough sound classification are presented in Table VI. With an accuracy of 0.6805, $F_1$ of 0.6824, and AUC of 0.7476 the best results using the RF classifier are on the DL feature set. As such, the DL features outperformed the handcrafted features with RF classification and feature combination did not lead to improvement in the classification results. The feature dimension of the DL feature set is the smallest among the three feature sets and, in this case, the RF classifier seems to generalize well on the small feature set.

Feature combination is once again seen to be ineffective in improving the classification results with the SVM classifier. However, unlike the RF classifier, the highest accuracy of 0.7278, $F_1$ of 0.7262, and AUC of 0.7692 are achieved on the handcrafted feature set, outperforming the DL features. Also, the best results using the SVM classifier are higher than the best results using the RF classifier.

When utilizing the MLP classifier with the handcrafted and combined feature sets, we observe improvements in all performance metrics. With the handcrafted feature set, the MLP classifier attains an accuracy of 0.7396, marking a relative enhancement of 11.60% compared to RF and a 1.62% improvement over SVM. The $F_1$ score reaches 0.7381, showing a relative increase of 11.38% over RF and 1.64% over SVM, while the AUC is 0.7771, indicating a relative improvement of 7.16% over RF and 1.03% over SVM. On the combined feature set, the MLP classifier achieves an accuracy of 0.7633, signifying a relative improvement of 15.18% over RF and 7.49% over SVM. The $F_1$ score is 0.7619, exhibiting a relative increase of 14.28% over RF and 7.29% over SVM, and the AUC reaches 0.7994, with a relative improvement of 8.48% over RF and 4.41% over SVM. Consequently, the MLP classifier surpasses the RF and SVM classifiers in performance, and the best classification results are attained when employing the combined feature set. Also, the classification results of the MLP improves with feature combination and, unlike the RF classifier, the MLP classifier is better able to learn on the combined feature set.

Next, the MLP learnings, in this case the activations from the second ReLU layer of the MLP network, for the pneumonia and non-pneumonia cough sound features in the test set are investigated using $t$-SNE. The $t$-SNE visualization, Fig. 8, shows that cough features from the pneumonia and non-pneumonia classes form visible clusters. This implies that the MLP network understands the cough sound features and its classes and is able to differentiate them. Furthermore, in Fig. 9, we present box plots for the most significant feature in both the handcrafted feature set and the deep learning feature set, which was identified based on the lowest $p$-value using the $t$-test. Notably, VGGish feature embedding 62 emerges as the most significant feature, followed by the standard deviation of the 11$^{\text{th}}$ mel-frequency cepstral coefficient. MFCCs represent different aspects of the spectral characteristics of the cough sound, with the lower coefficients generally capturing the overall shape and envelope of the spectrum and the higher coefficients capturing the finer spectral details and rapid changes in the spectrum. Our analysis shows that finer spectral characteristics in cough sounds are important in distinguishing pneumonia from other acute respiratory diseases.

MFCCs are one of the most commonly used features in audio classification tasks. However, the time derivatives of these static parameters were originally proposed for speech classification tasks and their contribution in cough sound classification hasn't been evaluated. In Fig. 10, we plot the sensitivity, specificity, and accuracy of MFCCs and its first and second derivatives, evaluated separately, for pneumonia

TABLE VI
COUGH SOUND CLASSIFICATION RESULTS INTO PNEUMONIA AND NON-PNEUMONIA

| Feature Set | Feature Selection Method | Classifier | Classification Results | | | | |
|---|---|---|---|---|---|---|---|
| | | | Sensitivity | Specificity | Accuracy | $F_1$ | AUC |
| Handcrafted features | None | RF | 0.7179 | 0.6154 | 0.6627 | 0.6627 | 0.7252 |
| DL features | T-Test | RF | 0.7436 | 0.6264 | 0.6805 | 0.6824 | 0.7476 |
| Handcrafted + DL features | T-Test | RF | 0.7308 | 0.6044 | 0.6627 | 0.6667 | 0.7369 |
| Handcrafted features | None | SVM | 0.7821 | 0.6813 | 0.7278 | 0.7262 | 0.7692 |
| DL features | Elastic Net | SVM | 0.7051 | 0.6044 | 0.6509 | 0.6509 | 0.7416 |
| Handcrafted + DL features | None | SVM | 0.7692 | 0.6593 | 0.7101 | 0.7101 | 0.7656 |
| Handcrafted features | T-Test | MLP | 0.7949 | 0.6923 | 0.7396 | 0.7381 | 0.7771 |
| DL features | None | MLP | 0.7179 | 0.6154 | 0.6627 | 0.6627 | 0.6844 |
| Handcrafted + DL features | T-Test | MLP | **0.8205** | **0.7143** | **0.7633** | **0.7619** | **0.7994** |

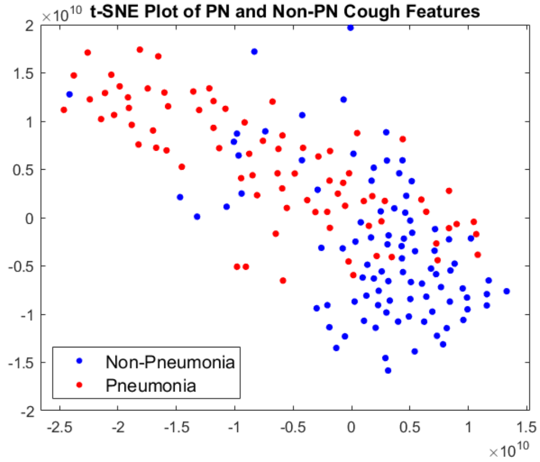

Fig. 8. *t*-SNE visualization of MLP network activations for pneumonia (PN) vs non-pneumonia (non-PN) cough classification.

vs non-pneumonia cough sound classification using MLP. MFCCs on their own achieve an accuracy of 0.7101 followed by the first derivatives with an accuracy of 0.6627 and the second derivatives with an accuracy of 0.6331. As such, while MFCCs are more discriminative of pneumonia and non-pneumonia cough sounds then its derivatives, the derivatives of MFCCs also carry discriminative characteristics and this leads to improvement in classification performance when these features are combined (Table VI). We also note that, unlike other non-speech sounds, cough shares some similarities with speech in terms of the physiological systems involved in the generation process.

## IV. Discussion and Conclusion

This work proposes a fully automated method of cough sound analysis in classifying pneumonia and non-pneumonia in children. The proposed method involves cough denoising, cough segmentation, and cough sound classification. With a sensitivity of 0.8205, specificity of 0.7143, accuracy of 0.7633, $F_1$ of 0.7619, and AUC of 0.7994, these are the best results of all the methods considered in this work. In addition, these results are only marginally lower than what could be achieved when the cough sounds are manually segmented [22]. The combination of sensitivity and specificity achieved in our work is moderately to significantly better than what has been reported using the WHO clinical algorithm to identify pneumonia, as summarized in previous studies [6], [7].

In pneumonia vs non-pneumonia cough sound classification, we studied two variations over earlier works. These are based on neural networks, with the incorporation of feature embeddings from a pretrained convolutional neural network

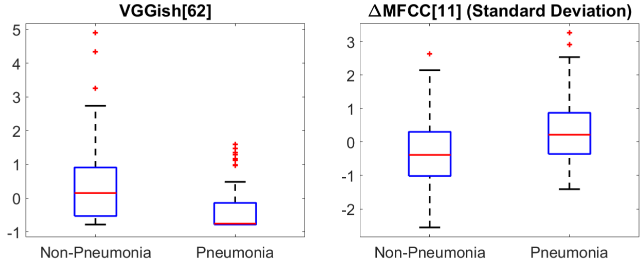

Fig. 9. Box plots displaying the most significant feature (with the lowest *p*-value) from each feature set.

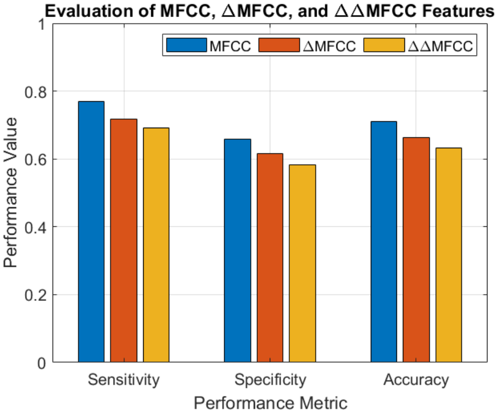

Fig. 10. Sensitivity, specificity, and accuracy of MFCCs and its first and second derivatives in pneumonia vs non-pneumonia cough sound classification.

and the use of MLP for classification. On the combined feature set, where we achieve the best results using MLP, we observe an improvement in accuracy, $F_1$, and AUC in the range of 4.41% to 15.18% over the corresponding results using RF and SVM. In addition, the inclusion of the deep learning features provides an improvement in accuracy, $F_1$, and AUC in the range of 2.87% to 3.22% over the results using the handcrafted features only. As such, the use of neural network methods helps us achieve better classification results in detecting pneumonia from non-pneumonia using cough sounds. In addition, the advantages of neural networks over conventional methods in cough sound denoising and segmentation have been demonstrated in earlier studies [24], [29].

The proposed method of detecting childhood pneumonia using automated cough sound analysis can be implemented on smartphones and possibly find several applications. Errors in diagnosing respiratory diseases are common in healthcare settings [47], [48], [49]. The proposed method can provide clinicians an additional tool to aid their decision making, such as in triaging and primary healthcare. Deaths due to childhood pneumonia are highest in remote communities where such a tool would also be useful for screening for the disease. In addition, the viruses and bacteria that cause pneumonia can be contagious and visiting a physician can lead to other children getting infected. The utilization of virtual healthcare has seen a significant increase globally during the COVID-19 pandemic, and this trend is widely anticipated to continue in the future [50]. The proposed method can be integrated into telehealth platforms to provide objective cough sound evaluation to the physician during telehealth consultation.

However, our study does have certain constraints. Although our dataset includes a larger number of subjects compared to similar studies like [6], [18], it's important to note that the non-pneumonia group predominantly consists of individuals with bronchitis. In our future research, we intend to further assess our approach by incorporating a more diverse set of cough recordings obtained from various acute pediatric respiratory conditions. Also, to mitigate the losses due to MP3 conversion, we analyzed the frequency content up to 8 kHz but the power spectrum of pneumonia cough extends up to 20 kHz [6]. In the future, we plan to use uncompressed audio formats or formats with lossless compression to prevent loss of spectral content due to lossy compression. In addition, in our work, there is an average of less than 3 coughs per subject. In contrast, the test dataset of two similar studies [6], [18] has 15 coughs per subject while the dataset of another two studies [19], [20] consists of 5 coughs per subject. Having more cough samples per subject can potentially provide better estimate of feature statistics, such as mean and standard deviation that we use in our work. As such, in the future we plan to experiment with greater number of cough samples per subject.

Furthermore, the proposed method has been only internally validated. Validation of cough sound-based respiratory disease detection algorithms on external or independent data can be challenging, with a drop in classification performance observed in diagnostic studies on acute pediatric respiratory diseases [19], [20] and COVID-19 [51], [52]. Similar shortcomings have been observed in external validation of AI prediction models in other medical applications as well, such as in sepsis prediction using electronic health record data [53] and image-based radiologic diagnosis [54]. Differences in classification performance from internal validation to external validation can be due to several reasons, including data heterogeneity [55] and variability in ground truthing [56]. It's yet to be determined how our proposed method will perform on an external dataset, such as from a different healthcare setting. In the future, we plan to perform such diagnostic studies using techniques such as data standardization [55] and using expert graders to reduce variability in ground truthing [56] to help mitigate the shortcomings of previous studies.

## References

[1] *Fact sheet: Pneumonia in children [Internet]*: World Health Organization, 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/pneumonia.

[2] D. Burgner and P. Richmond, "The burden of pneumonia in children: an Australian perspective," *Paediatric Respiratory Reviews*, vol. 6, no. 2, pp. 94–100, 2005.

[3] T. Kovesi, "Respiratory disease in Canadian First Nations and Inuit children," *Paediatrics & Child Health*, vol. 17, no. 1, pp. 376–380, 2012.

[4] K. Kallander, D. H. Burgess, and S. A. Qazi, "Early identification and treatment of pneumonia: a call to action," *The Lancet Global Health*, vol. 4, no. 1, pp. e12–e13, 2016.

[5] World Health Organization, *Pocket Book of Hospital Care for Children: Guidelines for the Management of Common Childhood Illnesses*, $2^{\text{nd}}$ ed. Geneva: WHO Press, 2013.

[6] U. R. Abeyratne, V. Swarnkar, A. Setyati, and R. Triasih, "Cough sound analysis can rapidly diagnose childhood pneumonia," *Annals of Biomedical Engineering*, vol. 41, no. 11, pp. 2448–2462, 2013.

[7] P. P. Moschovis *et al.*, "The diagnosis of respiratory disease in children using a phone-based cough and symptom analysis algorithm: The smartphone recordings of cough sounds 2 (SMARTCOUGH-C 2) trial design," *Contemporary Clinical Trials*, vol. 101, p. 106278, 2021.

[8] K. E. Fleming-Dutra *et al.*, "Prevalence of inappropriate antibiotic prescriptions among US ambulatory care visits, 2010-2011," *Journal of the American Medical Association*, vol. 315, no. 17, pp. 1864–1873, 2016.

[9] S. B. Meropol, A. R. Localio, and J. P. Metlay, "Risks and benefits associated with antibiotic use for acute respiratory infections: a cohort study," *The Annals of Family Medicine*, vol. 11, no. 2, pp. 165–172, 2013.

[10] C. Biagi *et al.*, "Lung ultrasound for the diagnosis of pneumonia in children with acute bronchiolitis," *BMC Pulmonary Medicine*, vol. 18, no. 1, p. 191, 2018.

[11] Y. Feng *et al.*, "Deep supervised domain adaptation for pneumonia diagnosis from chest x-ray images," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 3, pp. 1080–1090, 2022.

[12] A. B. Chang, "The physiology of cough," *Paediatric Respiratory Reviews*, vol. 7, no. 1, pp. 2–8, 2006.

[13] J. Korpáš, J. Sadloňová, and M. Vrabec, "Analysis of the cough sound: an overview," *Pulmonary Pharmacology*, vol. 9, no. 5, pp. 261–268, 1996.

[14] R. V. Sharan, U. R. Abeyratne, V. R. Swarnkar, and P. Porter, "Automatic croup diagnosis using cough sound recognition," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 2, pp. 485–495, 2019.

[15] 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, p. 102722, 2021.

[16] Y. Chung, J. Jin, H. I. Jo, *et al.*, "Diagnosis of pneumonia by cough sounds analyzed with statistical features and AI," *Sensors* vol. 21, no. 21, p. 7036, 2021.

[17] M. Binnekamp, K. J. van Stralen, L. den Boer, and M. A. van Houten, "Typical RSV cough: myth or reality? A diagnostic accuracy study," *European Journal of Pediatrics*, vol. 180, no. 1, pp. 57–62, 2021.

[18] K. Kosasih, U. R. Abeyratne, V. Swarnkar, and R. Triasih, “Wavelet augmented cough analysis for rapid childhood pneumonia diagnosis,” *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 4, pp. 1185–1194, 2015.

[19] P. Porter *et al.*, “A prospective multicentre study testing the diagnostic accuracy of an automated cough sound centred analytic system for the identification of common respiratory disorders in children,” *Respiratory Research*, vol. 20, no. 1, p. 81, 2019.

[20] P. P. Moschovis *et al.*, “A cough analysis smartphone application for diagnosis of acute respiratory illnesses in children,” in *Proceedings of the American Thoracic Society International Conference*, Dallas, Texas, 2019, p. A1181.

[21] S. Liao, C. Song, X. Wang, and Y. Wang, “A classification framework for identifying bronchitis and pneumonia in children based on a small-scale cough sounds dataset,” *PLOS ONE*, vol. 17, no. 10, p. e0275479, 2022.

[22] R. V. Sharan, K. Qian, and Y. Yamamoto, “Detecting childhood pneumonia using handcrafted and deep learning cough sound features and multilayer perceptron,” in *Proceedings of the 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Sydney, Australia, 2023, pp. 1–4.

[23] R. V. Sharan, “Productive and non-productive cough classification using biologically inspired techniques,” *IEEE Access*, vol. 10, pp. 133958–133968, 2022.

[24] L. Jose, S. Berkovsky, H. Xiong, C. Mascolo, and R. V. Sharan, “Denoising cough sound recordings using neural networks,” in *Proceedings of the 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Sydney, Australia, 2023, pp. 1–4.

[25] D. Liu, P. Smaragdis, and M. Kim, “Experiments on deep learning for speech denoising,” in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Singapore, 2014, pp. 2685–2689.

[26] S. R. Park and J. W. Lee, “A fully convolutional neural network for speech enhancement,” in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Stockholm, Sweden, 2017, pp. 1993–1997.

[27] M. Ravanelli and Y. Bengio, “Speaker recognition from raw waveform with SincNet,” in *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, Athens, Greece, 2018, pp. 1021–1028.

[28] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, “On the properties of neural machine translation: encoder–decoder approaches,” in *Proceedings of the Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, Doha, Qatar, 2014, pp. 103–111.

[29] R. V. Sharan, “Cough sound detection from raw waveform using SincNet and bidirectional GRU,” *Biomedical Signal Processing and Control*, vol. 82, p. 104580, 2023.

[30] Y. Hu and Z. F. Jiang, *Zhu Fu Tang Practical Pediatrics*, 8$^{\text{th}}$ ed. Beijing: People’s Health Publishing House, 2015.

[31] N. M. Papadakis, I. Aroni, and G. E. Stavroulakis, “Effectiveness of MP3 coding depends on the music genre: evaluation using semantic differential scales,” *Acoustics*, vol. 4, no. 3, pp. 704–719, 2022.

[32] L. Orlandic, T. Teijeiro, and D. Atienza, “The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms,” *Scientific Data*, vol. 8, no. 1, p. 156, 2021.

[33] A. Varga and H. J. M. Steeneken, “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[34] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” *arXiv preprint arXiv:1502.03167*, 2015.

[35] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010, pp. 807–814.

[36] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[37] B. R. Glasberg and B. C. Moore, “Derivation of auditory filter shapes from notched-noise data,” *Hearing Research*, vol. 47, no. 1–2, pp. 103–138, 1990.

[38] A. L. Maas, A. Y. Hannun, and A. Y. Ng, “Rectifier nonlinearities improve neural network acoustic models,” in *Proceedings of the International Conference on Machine Learning*, Atlanta, Georgia, USA, 2013,

[39] C. M. Bishop, *Pattern Recognition and Machine Learning.* New York: Springer, 2006.

[40] 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.

[41] S. Young *et al.*, *The HTK book (for HTK version 3.4).* Cambridge University Engineering Department, 2009.

[42] S. Hershey *et al.*, “CNN architectures for large-scale audio classification,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, 2017, pp. 131–135.

[43] 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.

[44] A. Mesaros, T. Heittola, and T. Virtanen, “Metrics for polyphonic sound event detection,” *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.

[45] A. B. Chang, “Cough: are children really different to adults?,” *Cough*, vol. 1, no. 1, p. 7, 2005.

[46] L. van der Maaten and G. Hinton, “Visualizing data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

[47] P. Porter *et al.*, “Diagnostic errors are common in acute pediatric respiratory disease: A prospective, single-blinded multicenter diagnostic accuracy study in Australian emergency departments,” *Frontiers in Pediatrics*, vol. 9, p. 736018, 2021.

[48] M. Haddad *et al.*, “Errors in diagnosing infectious diseases: A physician survey,” *Frontiers in Medicine*, vol. 8, p. 779454, 2021.

[49] I. F. Jørgensen and S. Brunak, “Time-ordered comorbidity correlations identify patients at risk of mis- and overdiagnosis,” *npj Digital Medicine*, vol. 4, no. 1, p. 12, 2021.

[50] P. Webster, “Virtual health care in the era of COVID-19,” *The Lancet*, vol. 395, no. 10231, pp. 1180–1181, 2020.

[51] ResApp Health Limited, “ResApp announces positive results for a new novel smartphone-based COVID-19 screening test,” Brisbane, Australia, 2022. https://www.resapphealth.com.au/wp-content/uploads/2022/03/2358427.pdf (Accessed 1 August 2023).

[52] ResApp Health Limited, “Results from data confirmation study,” Brisbane, Australia, 2022. https://www.resapphealth.com.au/wp-content/uploads/2022/06/2396080.pdf (Accessed 1 August 2023).

[53] A. Wong *et al.*, “External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients,” *JAMA Internal Medicine*, vol. 181, no. 8, pp. 1065–1070, 2021.

[54] A. C. Yu, B. Mohajer, and J. Eng, “External validation of deep learning algorithms for radiologic diagnosis: A systematic review,” *Radiology: Artificial Intelligence*, vol. 4, no. 3, p. e210064, 2022.

[55] J. He, S. L. Baxter, J. Xu *et al.*, “The practical implementation of artificial intelligence technologies in medicine,” *Nature Medicine*, vol. 25, pp. 30–36, 2019.

[56] P.-H. C. Chen, C. H. Mermel, and Y. Liu, “Evaluation of artificial intelligence on a reference standard based on subjective interpretation,” *The Lancet Digital Health*, vol. 3, no. 11, pp. e693–e695, 2021.