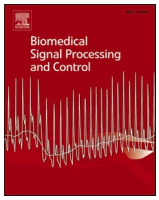
Biomedical Signal Processing and Control 82 (2023) 104580

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc



Cough sound detection from raw waveform using SincNet and bidirectional GRU

Roneel V. Sharan

Australian Institute of Health Innovation, Macquarie University, Sydney, NSW 2109, Australia

ARTICLE INFO

Keywords:
Convolutional neural network
Cough detection
Gammatone filters
Gated recurrent unit
Sinc functions

ABSTRACT

Background and objective: Cough is a common symptom of respiratory diseases and the sound of cough helps in understanding the condition of the respiratory system. Objective artificial intelligence driven cough sound evaluation has the potential to aid clinicians in diagnosing respiratory diseases. Automatic cough sound detection is an important step in performing objective cough sound analysis. Current methods in automatic cough sound detection involves various signal transformation and feature engineering steps which are not only complex, but can also lead to loss of signal characteristics and thereby suboptimal classification performance. This work aims to develop algorithms for robust cough sound detection directly from the audio recordings.

Methods: The proposed method utilizes SincNet, a one-dimensional convolutional neural network that uses sinc functions in the first convolutional layer to discover meaningful filters in the audio signal, and bidirectional gated recurrent unit, a type of recurrent neural network to learn the bidirectional temporal dependencies between the sequences in the audio signal. The filter parameters of the SincNet are initialized using the model of the human auditory filters. The proposed approach is evaluated on a manually annotated dataset of 400 audio recordings, containing more than 72,000 cough and non-cough frames.

Results: A validation accuracy of 0.9509 (AUC = 0.9903) and test accuracy of 0.9496 (AUC = 0.9866) is achieved in detecting cough and non-cough frames in the audio recordings using the proposed method.

Conclusion: The proposed cough detection approach forgoes the need for signal transformation and feature engineering and outperforms multiple baseline methods.

1. Introduction

Respiratory diseases are among the most common causes of illness and death worldwide [1,2]. Cough is a common symptom in respiratory diseases. Different respiratory diseases can affect the airways differently and, therefore, can cause variations in cough sounds, such as productive (wet) and non-productive (dry) [3], the distinctive barking cough of croup [4], and whooping cough [5]. While the cough of COVID-19 is not as well understood, several research groups around the world are studying it.

Automatic cough sound detection is an important part of cough sound analysis algorithms [4]. In earlier works, cough detection is a multistage process. In [6], various handcrafted features and time delay neural network are used for classification. A similar approach is also adopted in [4]. Handcrafted features are also used in [7], extracted using the openSMILE toolkit [8], for classification using random forest and gradient boosting, two ensemble classifiers.

A multistage cough sound detection approach is also employed in [9], but with deep learning classification. The audio signal is preprocessed by removing silence and low energy windows. The remaining time windows are then frequency transformed using short-time Fourier transform (STFT). Two deep learning approaches are experimented with. Firstly, the resulting time-frequency image-like representation is used for classification using two-dimensional convolutional neural networks (CNN). Secondly, they experiment with recurrent neural networks (RNN), to learn the dependencies between the frequency transformed frames.

However, deep learning on the frequency transformed signals presents less learnable parameters for the deep learning model. In addition, the multistage processes in cough sound detection employed in these earlier works can be time consuming and complex. The various signal transformations can also lead to loss of signal characteristics and thereby compromise the accuracy in cough detection.

It is possible to feed the raw audio signal frames as direct input to standard CNNs, as seen in speech recognition [10]. The first convolutional layer of such raw waveform-based CNNs is important as it deals

E-mail address: roneel.sharan@mq.edu.au.

https://doi.org/10.1016/j.bspc.2023.104580
Received 12 July 2022; Received in revised form 5 November 2022; Accepted 9 January 2023

with high-dimensional waveform input and is more affected by the vanishing gradient problem, but the filters learned by the network take noisy and discordant shapes [11]. This leads to an inefficient representation of the sound signals. As such, while the raw audio signal segments can be fed directly to a standard CNN, it will only learn low-level cough sound representations from the waveforms.

This work aims to perform cough sound detection using raw audio signals; however, it proposes to introduce some constraints on the shape of the CNN filters in the input layer to help the CNN learn more meaningful filters. The proposed method is inspired by *SincNet* [11], a deep learning network architecture originally proposed for speaker recognition using raw waveform. Unlike the standard CNN where the filterbank characteristics depend on several parameters, the SincNet utilizes a set of parameterized sinc functions with which the raw waveform is convolved. The sinc functions implement bandpass filters and the cutoff frequencies are the only parameters learned. This helps the network learn high-level tunable parameters with broad impact on the design of the resulting filter and, therefore, discover more meaningful characteristics in the audio signal. SincNet has been shown to produce faster convergence of the network during training and performance gains over the standard CNN [11].

This work proposes two improvements to the original SincNet. Firstly, it proposes the use of gammatone filterbank for initializing the SincNet filter parameters. Gammatone filters are a widely used model of auditory filters and have the advantage of providing finer frequency characterization at low frequencies where important cough characteristics are located [4]. Gammatone filters can be considered an improvement over the triangular filters conventionally used in mel filters. Secondly, it proposes the use of gated recurrent units (GRU) [12]. GRU is a type of RNN, similar to long short-term memory (LSTM) networks, but with a smaller number of gates and parameters. A comparison of LSTM and GRU shows that GRU can outperform LSTM in convergence and in parameter updates and generalization. In particular, this work proposes the use of bidirectional GRU (BiGRU) that uses a finite sequence to predict the class of each cough or non-cough frame of a sequence of frames based on its past and future contexts. As such, while the SincNet will learn the spectral characteristics within the frame, this work will use BiGRU to learn the temporal dependencies between the frames. The proposed *SincNet-BiGRU* method is evaluated on a dataset of 400 manually annotated cough and non-cough audio recordings containing more than 72,000 cough and non-cough frames. The performance of the proposed method is compared against several baseline methods to demonstrate its effectiveness in detecting cough and non-cough sound signals.

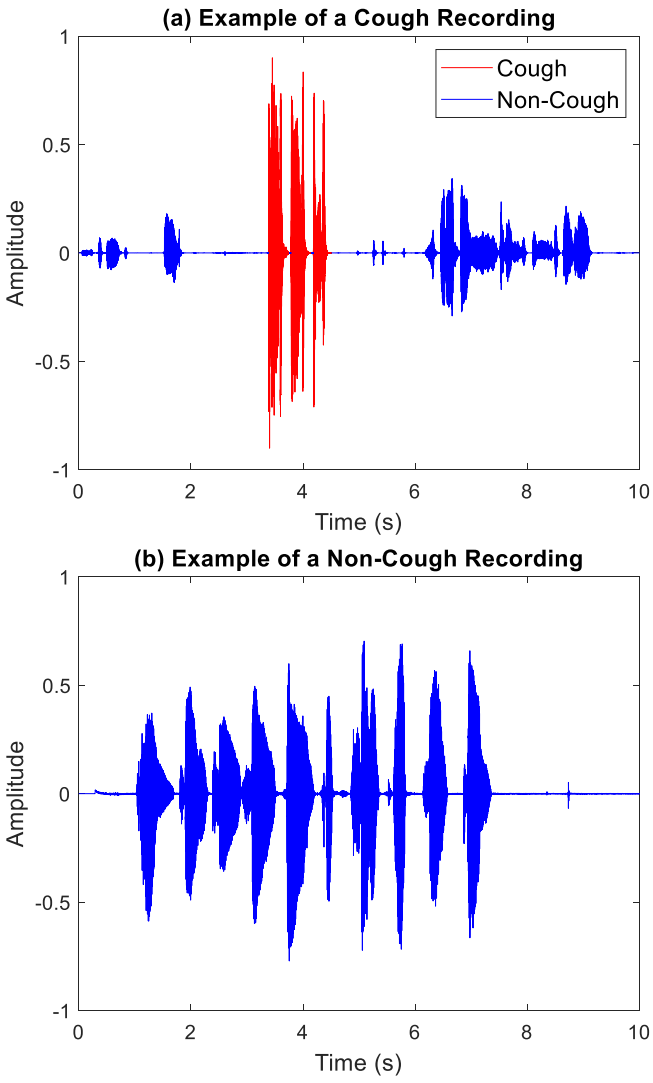


Fig. 1. Illustration of (a) a cough recording and (b) a non-cough recording. The cough recording comprises of both, cough and non-cough segments (speech, non-speech, and silence) while the non-cough recording comprises of non-cough sounds only. In this illustration, the subject is counting from 1 to 10 in the non-cough recording.

2. Method

2.1. Dataset

The dataset of cough sound recordings used in this work was collected as part of COVID-19 research [13]. The dataset has been crowdsourced from subjects from around the world, likely using different recording devices and recorded in different environments, making this a very diverse and challenging dataset. The original dataset has 27,550 recordings.

A fraction of the cough recordings has been annotated by up to four expert physicians in the original dataset. This work utilizes 200 audio recordings from this subset, referred here as the *cough sound recordings*. These recordings either contain cough sounds and silence or cough sounds, silence, and other non-cough sounds, such as speech, music, radio, television, etc. The cough sounds in these 200 recordings are manually segmented by the author for supervised cough sound detection. An example of a cough recording containing cough sounds, silence, and speech is illustrated in Fig. 1(a).

While it is important for a cough detection algorithm to be able to detect cough sounds, it is also important to reject non-cough sounds. The cough sound recordings selected for this work contain non-cough sounds as well but the amount of non-cough sounds that are non-silence is limited in these recordings. For this reason, an additional 200 audio recordings are manually annotated by the author from the overall dataset that do not contain any cough sounds, referred as *non-cough sound recordings*. An example of a non-cough sound recording is given in Fig. 1(b) where the subject is counting from 1 to 10 instead of coughing.

The audio recordings are sampled at 16 kHz and in waveform audio file format (WAV). An overview of the final dataset used in this work is provided in Table 1. The overall dataset of 400 recordings is divided into a training and validation set of 300 recordings (150 cough recordings and 150 non-cough recordings) and a test set of 100 recordings (50 cough recordings and 50 non-cough recordings). The average duration of the training and validation recordings is 8.90 s and 8.28 s for the test recordings. The cough recordings in the training and validation set contain a total of 683 cough sounds while the cough recordings in the test set contain 232 cough sounds. Using a window size of 64 ms, similar to [9], with 25% overlap between frames results in 55,332 cough and non-cough frames in the training and validation set and 17,169 cough and non-cough frames in the test set. The average age of the subjects in the training and validation set is 36.09 and 39.73 in the test set. While

Table 1
Overview of the training and validation, and test datasets used in this work.

		Training and validation recordings			Test recordings		
		Cough	Non-Cough	All	Cough	Non-Cough	All
Number of recordings		150	150	300	50	50	100
Number of coughs in recordings		683	0	683	232	0	232
Average recording duration (seconds)		8.76	9.04	8.90	8.48	8.09	8.28
Number of frames (64 ms, 25% overlap)		27,229	28,103	55,332	8,785	8,384	17,169
Average age (years)		35.57	38.00	36.09	38.45	41.96	39.73
Gender	Male	92	22	114	31	13	44
	Female	40	22	62	13	16	29
	Other/Unknown	18	106	124	6	21	27

not all subjects reported their age, it was aimed to match the age of the subjects based on the available data. The gender of the subjects is also summarized in Table 1. Based on the analysis in [13], the SNR of the recordings used in this work is estimated to be 13.66 ± 13.58 dB. Detailed description of the full dataset, including demographics and clinical data, can be found in [13].

2.2. Deep learning network

An overview of the proposed deep learning network for raw waveform based cough sound detection is shown in Fig. 2. The audio signals are converted into a sequence of frames (sequences of 64 ms (1024 points) with 25% overlap between sequences) at the sequence input layer. The number of sequences depends on the length of the signal. The sequence folding layer facilitates the convolutional operations, in the SincNet, on time steps of the audio signal sequences independently. A sequence unfolding layer followed by a flatten layer restore the sequence structure and reshape the output to vector sequences. The resulting vector sequences are classified using a bidirectional GRU which learns bidirectional long-term dependencies between the time steps of the sequence data. The final layers of the network include a fully connected layer, softmax layer [14], and classification layer.

2.2.1. SincNet

The SincNet comprises of 3 sets of convolutional layers. The first layer performs sinc-based convolutions using 80 filters of length 251. The next two are standard convolutional layers using 60 filters of length 5. Each layer is followed by a batch normalization layer [15], a leaky rectified linear unit [16] with a scalar multiplier for negative inputs equal to 0.2, and a 1 × 3 max pooling layer. The stride for all convolutional and max pooling layers is 1. This is followed by three fully connected layers. Each fully connected layer has an output size of 256 and is followed by batch normalization and leaky rectified linear unit layers.

When using the raw time-domain cough or non-cough audio signal as input to a standard CNN, the first layer performs convolution between the input waveform and some finite impulse response filter, given as [17]

$$y[n] = x[n] * h[n] = \sum_{l=0}^{L-1} x[l] \cdot h[n-l] \tag{1}$$

where $x[n]$ is a segment from a cough or non-cough audio signal, $h[n]$ is the filter of length L, and $y[n]$ is the filtered output. All L elements of the filter are learned from the data in the standard CNN.

In the SincNet, convolution is performed with a predefined function g that depends on only few learnable parameters θ, given as

$$y[n] = x[n] * g[n, \theta] \tag{2}$$

and the frequency response of g is a rectangular function [11]. In frequency-domain, the magnitude response of a bandpass filter is essentially a difference of two such filters given as

$$G(f) = \text{rect}\left(\frac{f}{2f_2}\right) - \text{rect}\left(\frac{f}{2f_1}\right) \tag{3}$$

where f_1 and f_2 are the lower and upper cutoff frequency of the bandpass filter and *rect*(·) is the rectangular function [18]. In time-domain, this transforms to

$$g(n) = 2f_2\text{sinc}(2\pi f_2 n) - 2f_1\text{sinc}(2\pi f_1 n) \tag{4}$$

where the *sinc* function is given as $\text{sinc}(x) = \sin(x)/x$.

The cutoff frequencies are initialized in the range $[0, f_S/2]$, where f_S is the sampling frequency of the audio signal. In this work, it is performed using the equivalent rectangular bandwidth [19], a psychoacoustic measure of the width of the human auditory filters, described as

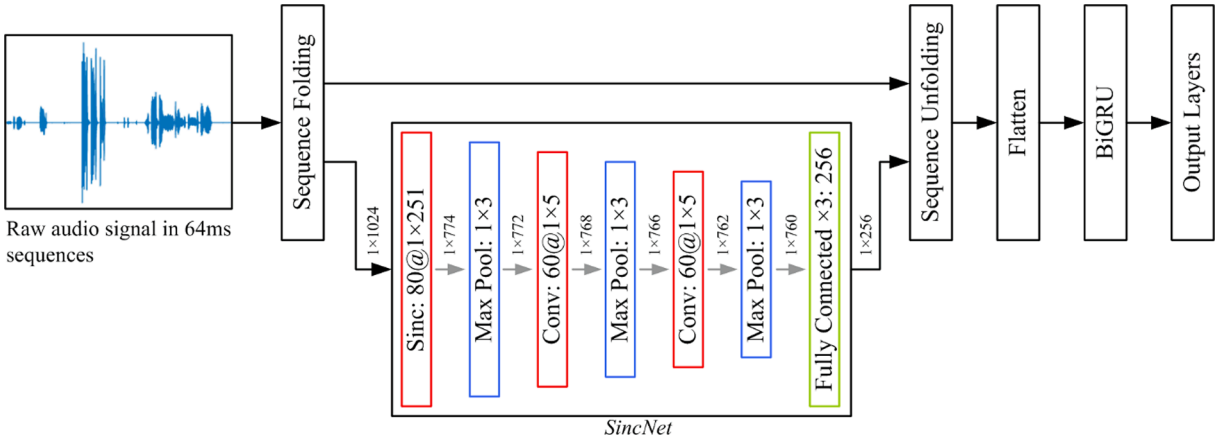


Fig. 2. An overview of the proposed method in cough sound detection. The method is based on *sequence-to-sequence* classification. It uses SincNet to learn intra-frame characteristics from the raw signal and bidirectional gated recurrent unit to learn the temporal dependencies between the sequence of frames.

$$ERB(f) = 24.7\left(\frac{4.37f}{1000} + 1\right) \tag{5}$$

where f is the center frequency of the filter in Hz and $ERB(f)$ is the bandwidth in Hz. The relationship between the number of ERBs to frequency is obtained by integrating the reciprocal of (5). The number of ERBs is then obtained as [19]

$$E = 21.4\log_{10}\left(\frac{4.37f}{1000} + 1\right). \tag{6}$$

The SincNet layer then aims to learn better parameters for these bandpass filters within the neural network framework. SincNet offers various advantages over the standard CNN, such as fast convergence, less number of parameters, computational efficiency, and interpretability of the convolutional layer [11].

2.2.2. BiGRU

The GRU learns dependencies between time steps in sequence data. At time step t, the hidden state of the GRU layer contains the output of the layer for this time step. Information is added or removed from the state at each time step using gates. The hidden state of the GRU layer is controlled by the reset gate r, update gate z, and the candidate state $\hat{h}$. The input weights W, recurrent weights R, and bias b, are the learnable weights of the GRU. The input and recurrent weight matrices are concatenations of each component (r, z, and $\hat{h}$) and the bias vector depends on whether the reset gate is applied to hidden state before or after matrix multiplication.

The hidden state at time step t is given as

$$\boldsymbol{h}_t = (1 - z_t) \odot \widehat{\mathbf{h}}_t + z_t \odot \boldsymbol{h}_{t-1} \tag{7}$$

and the components (r, z, and $\hat{h}$) at time step t depend on the mode of the reset gate. This work makes use of two GRUs to learn sequence information in both directions, backwards and forwards.

2.2.3. Weighted cross-entropy loss

The number of cough frames in the dataset is much less than the number of non-cough frames. During training, this class imbalance is accounted for using a weighted cross-entropy loss in the classification layer as

$$L = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{K} w_i t_{ni} \ln y_{ni} \tag{8}$$

where N is the number of samples, K the number of classes, w the class weights, t represents the training targets, and y the prediction scores of the softmax function.

2.2.4. Network training

The adaptive moment estimation optimization algorithm [20] is used as the solver for the training network and the training parameters are tuned using a simple grid search. The final setting for the initial learn rate is 0.0001, mini batch size is 4, and maximum number of epochs is 10. The network was implemented in MATLAB R2022a and trained using a single NVIDIA V100 Tensor Core GPU. The training stops after the maximum number of epochs is reached.

2.3. Evaluation metrics

The evaluation of the proposed cough detection method is based on frame classification, similar to [9]. The performance is measured using sensitivity, specificity, and accuracy, where sensitivity is the fraction of cough frames that are correctly classified, specificity is the fraction of non-cough frames that are correctly classified, and accuracy is the fraction of all frames (cough and non-cough) that are correctly classified. The area under the curve (AUC) of the receiver operating characteristic (ROC) curve is also used, as a single measure of classification performance. In addition, F-score and the equal error rate (EER) are used as evaluation metrics. F-score (or F_1 score) is the harmonic mean of the precision and recall, where precision is the number of cough frames that are correctly classified divided by the number of all frames classified as cough and recall is the same as sensitivity. For calculating the sensitivity, specificity, accuracy, and the F-score, the optimal threshold on the ROC curve is determined as the point on the ROC curve that minimizes the distance to the point (0,1). The EER is the value at which the false negative rate and the false positive rate are equal. For sensitivity, specificity, accuracy, AUC, and the F-score, a value of 1 indicates ideal performance. The ideal value for EER is 0.

3. Results

3.1. Experimental setup

Cough and non-cough frame classification is performed using three different techniques: (i) using handcrafted features and conventional classifiers, (ii) using convolutional neural networks, and (iii) using bidirectional GRU. The performance of the methods is evaluated in stratified 10-fold cross-validation on the training and validation set of 300 recordings. That is, 30 recordings (15 cough recordings and 15 non-cough recordings) are used for validation and the remaining 270 recordings (135 cough recordings and 135 non-cough recordings) for training in each fold. Cross-validation is used to tune the network parameters, described in Section 2.2.4. The network parameters are then fixed and the network is trained on all the 300 recordings from the training and validation set and evaluated on the unseen test set of 100 recordings. Cross-validation and test results are presented for all classification methods considered in this work.

3.2. Results using MFCCs and baseline classifiers

The results using mel-frequency cepstral coefficients (MFCCs) and baseline classifiers are presented first. MFCCs are a commonly used feature in audio classification tasks, including cough detection as seen in [4,6,7,9]. As in [9], 13 MFCCs are extracted in each frame. Additionally, the first and second derivatives of the coefficients are computed [21], resulting in a 39 dimensional feature vector in each frame. Four baseline classifiers are used for classification of the 39-dimensional MFCC feature set. These are logistic regression (LR), naive Bayes (NB), random forest (RF), and support vector machine (SVM).

The cross-validation results for cough and non-cough classification using MFCCs and the four baseline classifiers are given in Table 2. An accuracy of 0.8361 (AUC = 0.9187) is achieved using the LR classifier. This reduces to an accuracy of 0.7473 (AUC = 0.8390) using NB, which has the lowest classification accuracy and AUC of the four baseline classifiers. Using the RF classifier results in an accuracy of 0.8583 (AUC = 0.9393). With an accuracy of 0.8698 (AUC = 0.9424), SVM yields marginally better classification performance than RF and the highest accuracy and AUC of the four baseline classifiers.

The test results for cough and non-cough classification using MFCCs and the four baseline classifiers is presented in Table 3. The results show a similar trend to the cross-validation results and the highest accuracy of 0.8678 and AUC of 0.9371 is once again achieved using SVM. The relative change in the AUC is within 1% for all classifiers except NB which has possibly benefited from the additional training data.

3.3. Results using convolutional neural networks

The classification results using CNNs are presented next using the same cross-validation approach as the baseline methods in Section 3.2. Two types of CNNs are considered: 2-D CNN and 1-D CNN. The 2-D CNN classification approach is based on [9] and forms another baseline method. Each frame is converted to a 64 × 16 spectrogram image using

Table 2
Cross-validation results for cough and non-cough sequences using MFCC and various baseline classifiers.

Feature/Input	Classifier	Sensitivity	Specificity	Accuracy	AUC	F1	EER
MFCC	LR	0.8280	0.8371	0.8361	0.9187	0.5259	0.1691
MFCC	NB	0.8393	0.7360	0.7473	0.8390	0.4217	0.2304
MFCC	RF	**0.8607**	0.8580	0.8583	0.9393	0.5714	0.1409
MFCC	SVM	0.8602	**0.8709**	**0.8698**	**0.9424**	**0.5919**	**0.1353**

Table 3
Test results for cough and non-cough sequences using MFCC and various baseline classifiers.

Feature/Input	Classifier	Sensitivity	Specificity	Accuracy	AUC	F1	EER
MFCC	LR	0.8694	0.8160	0.8239	0.9201	0.5932	0.1626
MFCC	NB	0.8781	0.7694	0.7855	0.8698	0.5473	0.2005
MFCC	RF	**0.8793**	0.8295	0.8369	0.9308	0.6141	0.1497
MFCC	SVM	0.8592	**0.8693**	**0.8678**	**0.9371**	**0.6574**	**0.1362**

STFT. This spectrogram image forms input to a 2-D CNN. The architecture of the 2-D CNN is as described in [9]. The 1-D CNN is the SincNet, the input to which are raw audio sequences. While in the original SincNet the frequencies and bandwidth of the sinc functions are initialized as equally spaced on the mel scale, this work also experiments with the ERB scale as described in Section 2.2.1.

The cross-validation results for cough and non-cough classification using CNNs are given in Table 4. The Spectrogram + CNN approach of [9] yields an accuracy of 0.8818 (AUC = 0.9521). This improves to an accuracy of 0.8829 (AUC = 0.9584) with the SincNet when the sinc functions are initialized using the mel scale. This further improves to an accuracy of 0.8906 (AUC = 0.9572) when the sinc functions in the SincNet are initialized using the ERB scale. This trend is also observed in the test results, Table 5, with the SincNet (ERB) classification method achieving the highest accuracy and AUC of 0.8721 and 0.9467, respectively. Also, the relative change in the AUC is within 1.5% for all three methods.

As such, the CNN methods produce higher classification accuracy and AUC than the conventional feature engineering and classification methods the results for which are given in Tables 2 and 3. In addition, the proposed SincNet, with sinc functions initialized using the ERB scale, produces the highest classification accuracy and AUC. While it performs only marginally better than the original SincNet [11], the method could be considered more explainable with the ERB. Also, it outperforms MFCC + SVM, the baseline classification method with the highest accuracy and AUC.

3.4. Results using bidirectional GRU

Finally, classification results using the BiGRU are presented. The BiGRU is used with the three different methods reported in Sections 3.2 and 3.3, i.e., with the MFCC feature set, the Spectrogram + CNN method, and the Raw Signal + SincNet method. The Raw Signal + SincNet-BiGRU architecture is illustrated in Fig. 2. It is essentially an extension of the SincNet classification method used in Section 3.3 but with the introduction of the BiGRU.

The cross-validation results for cough and non-cough classification using BiGRU are given in Table 6. The MFCC + BiGRU classification approach achieves an accuracy of 0.9413 (AUC = 0.9742). This is a relative improvement of 8.23% in accuracy and 3.37% in AUC over the MFCC + SVM classification method, which produced the highest accuracy and AUC using the baseline classifiers (Table 2).

The Spectrogram + CNN-BiGRU method achieves an accuracy of 0.9428 (AUC = 0.9845). This is a relative improvement of 6.93% in accuracy and 3.40% in AUC over the Spectrogram + CNN cross-validation results (Table 4).

An accuracy of 0.9504 and AUC of 0.9895 is achieved using the Raw Signal + SincNet-BiGRU classification method when the sinc functions are initialized using the mel scale. These are better than the Spectrogram + CNN-BiGRU classification method and a relative improvement of 7.64% in accuracy and 3.24% in AUC over the corresponding Raw Signal + SincNet classification method.

Finally, the proposed Raw Signal + SincNet-BiGRU classification method, where the sinc functions are initialized using the ERB scale, achieves an accuracy of 0.9509 (AUC = 0.9903). This is a relative improvement of 6.77% in accuracy and 3.45% in AUC over the corresponding Raw Signal + SincNet classification method. These are the highest accuracy and AUC of all the classification methods considered in this work.

In the multistage process in cough sound detection in [9], the silent frames are removed before classification. Since the method proposed in this work uses an end-to-end approach, silent and non-silent frames are all classified at once. It's possible that classifying silent frames as non-cough would be a relatively easy task which can overestimate the results in this work. In this regard, the performance of the Raw Signal + SincNet-BiGRU (ERB scale) classification method is reanalyzed after removing the silent frames and then recomputing the classification results. It results in a sensitivity of 0.9317, specificity of 0.9291, accuracy of 0.9297, AUC of 0.9799, F-score of 0.8573, and EER of 0.0698. As such, the proposed method demonstrates robustness in accurately detecting both silent and non-silent non-cough frames.

The test results in Table 7 also demonstrate a similar trend to the cross-validation results with the highest accuracy of 0.9496 and AUC of 0.9866 using the proposed Raw Signal + SincNet-BiGRU (ERB scale) classification method. In addition, the relative change in the AUC from cross-validation to test is within 0.5% indicating good generalizability of the classification networks.

Fig. 3 shows the frequency response of 6 of the 80 filters as learned by the proposed SincNet-BiGRU (ERB scale) network. These 6 filters have equally spaced indices in the interval of 1 to 80. The filter shapes

Table 4
Cross-validation results for cough and non-cough sequences using convolutional neural networks.

Feature/Input	Classifier	Sensitivity	Specificity	Accuracy	AUC	F1	EER
Spectrogram	CNN	0.8834	0.8815	0.8818	0.9521	0.6212	0.1179
Raw Signal	SincNet (Mel)	**0.9045**	0.8802	0.8829	**0.9584**	0.6290	0.1093
Raw Signal	SincNet (ERB)	0.8966	**0.8899**	**0.8906**	0.9572	**0.6429**	**0.1077**

Table 5
Test results for cough and non-cough sequences using convolutional neural networks.

Feature/Input	Classifier	Sensitivity	Specificity	Accuracy	AUC	F1	EER
Spectrogram	CNN	0.8919	0.8599	0.8646	0.9418	0.6605	0.1290
Raw Signal	SincNet (Mel)	0.8915	0.8614	0.8659	0.9442	0.6625	0.1279
Raw Signal	SincNet (ERB)	**0.8947**	**0.8682**	**0.8721**	**0.9467**	**0.6738**	**0.1231**

Table 6
Cross-validation results for cough and non-cough sequences using bidirectional gated recurrent unit.

Feature/Input	Classifier	Sensitivity	Specificity	Accuracy	AUC	F1	EER
MFCC	BiGRU	0.9363	0.9420	0.9413	0.9742	0.7780	0.0612
Spectrogram	CNN-BiGRU	0.9486	0.9421	0.9428	0.9845	0.7846	0.0550
Raw Signal	SincNet-BiGRU (Mel)	0.9565	0.9496	0.9504	0.9895	0.8089	0.0478
Raw Signal	SincNet-BiGRU (ERB)	**0.9590**	**0.9499**	**0.9509**	**0.9903**	**0.8109**	**0.0466**

Table 7
Test results for cough and non-cough sequences using bidirectional gated recurrent unit.

Feature/Input	Classifier	Sensitivity	Specificity	Accuracy	AUC	F1	EER
MFCC	BiGRU	0.9448	0.9135	0.9181	0.9740	0.7731	0.0758
Spectrogram	CNN-BiGRU	0.9471	0.9325	0.9346	0.9829	0.8106	0.0616
Raw Signal	SincNet-BiGRU (Mel)	**0.9546**	0.9451	0.9465	0.9854	0.8406	0.0518
Raw Signal	SincNet-BiGRU (ERB)	0.9491	**0.9497**	**0.9496**	**0.9866**	**0.8476**	**0.0509**

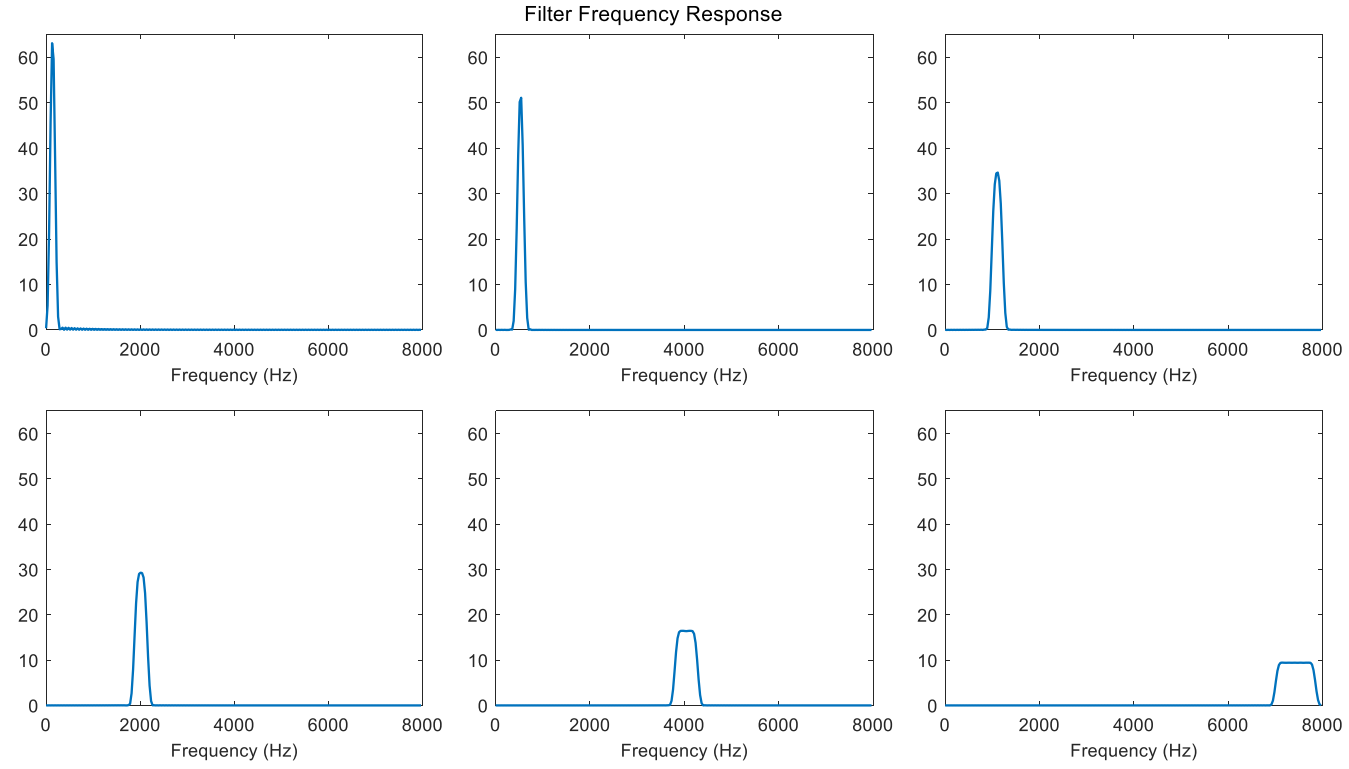


Fig. 3. Frequency response of 6 of the 80 filters learned by the proposed SincNet-BiGRU. The filters have equally spaced indices in the interval 1 to 80.

are narrow and sharp at low frequency values and the spacing of the filters is nonlinear. The spectrogram representations for the first cough and non-cough sound events from Fig. 1(a) and (b) are illustrated in Fig. 4(a) and (b), respectively. Both the spectrograms contain greater spectral content at low frequency values and the filters learned by the proposed network allow for finer frequency characterization at low frequency values.

Next, the SincNet-BiGRU predictions for the cough and non-cough recordings in the test set are investigated using *t*-distributed stochastic neighbor embedding (*t*-SNE) [22]. *t*-SNE maps high-dimensional data, network activations of the BiGRU, in this case, to two dimensions. The *t*-SNE visualization, Fig. 5, shows that data frames from the cough and non-cough classes form clearly visible clusters. This implies that the SincNet-BiGRU network understands the data frames and its classes and is able to differentiate them.

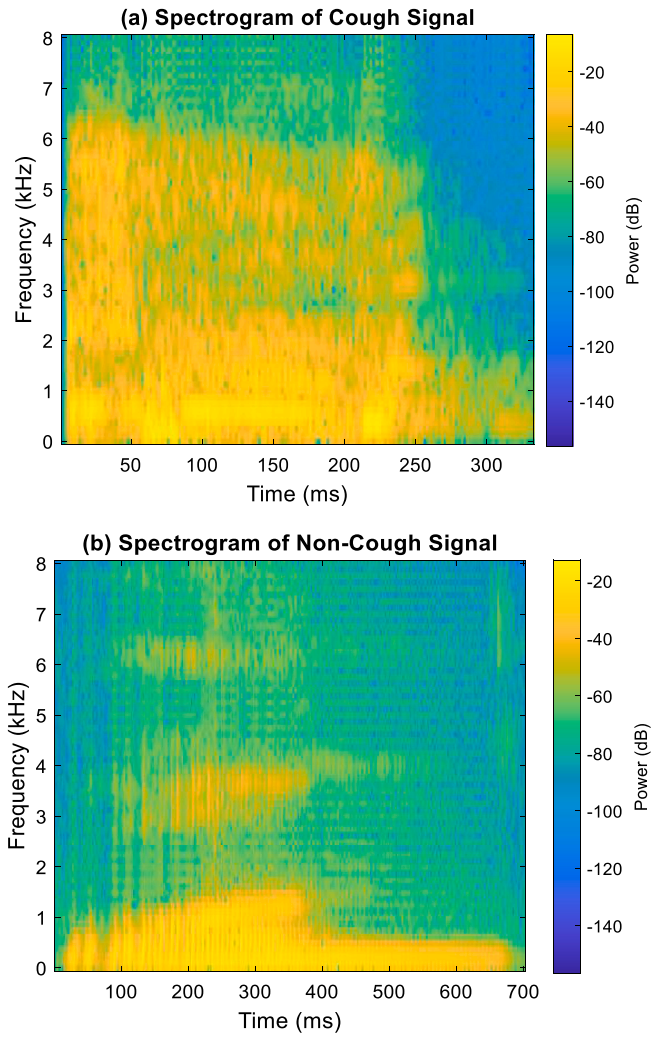


Fig. 4. Spectrogram representation of (a) a cough signal (first cough signal between 3.38 and 3.76 s from Fig. 1(a)) and (b) a non-cough signal (first speech signal between 1.02 and 1.72 s from Fig. 1(b)).

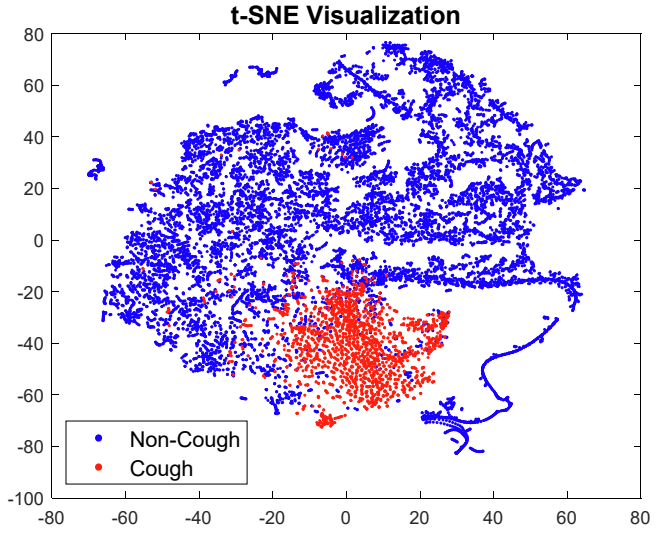


Fig. 5. *t*-SNE visualization of cough vs non-cough classification network activations.

4. Discussion

A method for detection of cough and non-cough frames using raw audio signal and deep learning techniques is presented in this paper. The deep learning method proposed in this work in based on the SincNet [11]. However, two improvements are proposed to the SincNet architecture. Firstly, the sinc filters are initialized on the ERB scale in the first layer. This first layer learns the spectral content of the cough and non-cough frames. As shown in Fig. 4, both signals exhibit greater spectral content at low frequencies compared to high frequencies. The ERB approximates the bandwidths of the filters in human hearing with greater frequency characterization at low frequencies (Fig. 3), making it ideal for this work. Secondly, the proposed SincNet (ERB) is combined with GRUs to learn the temporal dependencies between the frames. In particular, bidirectional GRUs are used to learn the temporal dependencies in both the forward and backward directions.

The performance of the proposed method is compared against MFCC features with four classifiers: LR, NB, RF, and SVM (results in Table 2 and 3). The Spectrogram-CNN [9] and SincNet (Mel) [11] methods are also used as baseline methods (results in Table 4 and 5). These baseline methods are implemented using the same dataset and experimental setup as the proposed method. The proposed Raw Signal + SincNet-BiGRU method outperforms the baseline methods which include handcrafted features, conventional classifiers, and CNN classification of time–frequency (spectrogram) image as proposed in [9].

In [23], a summary of different cough detection approaches is provided in Table 2, with an accuracy of 99.91% reported in [24] and 97% reported in [25]. On further analysis, it is noted that the work reported in [25] is not a cough detection task; it studies classification of COVID-19 coughs against healthy coughs and the coughs are *detected* using PRAAT software. Similarly, [24] looks at classification of productive cough, non-productive cough, and ambient sounds, with deep learning performed on data from only 8 recordings. This current study looks at automatic detection of cough sounds against non-cough sounds with training and validation on data from 300 recordings with an additional 100 recordings for testing only.

5. Conclusion

Cough is one of the most common presenting conditions in primary care [26]. There has been an increased uptake of virtual healthcare during COVID-19 and it is largely expected to continue [27]. Objective cough sound assessment using smartphone technology during virtual consultation can aid the physician in diagnosis of respiratory diseases and the proposed cough sound detection method can be useful in this regard as it is an important step of this process.

In addition, recently, many research groups around the world have been studying the cough sound of COVID-19 and several of these studies are reliant on crowdsourced data. Deep learning classification methods have outperformed various conventional classification methods in medical applications but require large datasets to train the deep learning model. While crowdsourced data can help in quickly collating large datasets for this purpose, crowdsourced data can be noisy and unreliable. Analysis of one COVID-19 dataset shows thousands of audio recordings possibly don't even contain cough sounds [13]. As such, the proposed method can be useful in detecting and segmenting cough sounds and also in discarding audio recordings that do not contain the sound of cough.

The dataset used in this work is crowdsourced which means varying experimental setup, such as the recordings are likely made using different devices with varying hardware and software characteristics, different environments presenting different background noise, and different microphone positioning. All of these can affect the recorded sound of cough. The proposed cough detection method achieves an accuracy of 0.9496 (AUC = 0.9866) on an unseen test dataset despite these challenges.

Funding

This work was supported by the Google Research Scholar Program and Macquarie University.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Roneel V. Sharan reports financial support was provided by Google Research.

Data availability

Data will be made available on request.

References

[1] Forum of International Respiratory Societies, The Global Impact of Respiratory Disease, 2nd ed., European Respiratory Society, Sheffield, 2017.
[2] GBD 2015 Mortality and Causes of Death Collaborators, Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: A systematic analysis for the Global Burden of Disease Study 2015, The Lancet, vol. 388, no. 10053, pp. 1459–1544, 2016.
[3] A. Murata, Y. Taniguchi, Y. Hashimoto, Y. Kaneko, Y. Takasaki, S. Kudoh, Discrimination of productive and non-productive cough by sound analysis, Intern. Med. 37 (9) (1998) 732–735.
[4] R.V. Sharan, U.R. Abeyratne, V.R. Swarnkar, P. Porter, Automatic croup diagnosis using cough sound recognition, IEEE Trans. Biomed. Eng. 66 (2) (2019) 485–495.
[5] R.V. Sharan, S. Berkovsky, D.F. Navarro, H. Xiong, A. Jaffe, Detecting pertussis in the pediatric population using respiratory sound events and CNN, Biomed. Signal Process. Control 68 (2021), 102722.
[6] Y.A. Amrulloh, U.R. Abeyratne, V. Swarnkar, R. Triasih, A. Setyati, Automatic cough segmentation from non-contact sound recordings in pediatric wards, Biomed. Signal Process. Control 21 (2015) 126–136.
[7] M.D. Kruizinga, et al., Development and technical validation of a smartphone-based pediatric cough detection algorithm, Pediatr. Pulmonol. 57 (3) (2022) 761–767.
[8] F. Eyben, M. Wöllmer, and B. Schuller, Opensmile: The munich versatile and fast open-source audio feature extractor, in Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 2010: Association for Computing Machinery, pp. 1459–1462.
[9] J. Amoh, K. Odame, Deep neural networks for identifying cough sounds, IEEE Trans. Biomed. Circ. Syst. 10 (5) (2016) 1003–1011.
[10] D. Palaz, M. Magimai-Doss, R. Collobert, Analysis of CNN-based speech recognition system using raw speech as input, in: Annual Conference of the International Speech Communication Association (INTERSPEECH), Dresden, Germany, 2015, pp. 11–15.
[11] M. Ravanelli, Y. Bengio, Speaker recognition from raw waveform with SincNet, in: IEEE Spoken Language Technology Workshop (SLT), Greece, Athens, 2018, pp. 1021–1028.
[12] K. Cho, B. van Merriënboer, D. Bahdanau, Y. Bengio, On the properties of neural machine translation: encoder–decoder approaches, in: Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 2014, pp. 103–111.
[13] L. Orlandic, T. Teijeiro, D. Atienza, The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms, Sci. Data 8 (1) (2021) 156.
[14] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, New York, 2006.
[15] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167, 2015.
[16] A.L. Maas, A.Y. Hannun, A.Y. Ng, Rectifier nonlinearities improve neural network acoustic models. International Conference on Machine Learning, Atlanta, Georgia, USA, 2013.
[17] L.R. Rabiner, R.W. Schafer, Theory and Applications of Digital Speech Processing, First ed., Prentice Hall, New Jersey, 2011.
[18] M. Owen, Practical Signal Processing, Cambridge University Press, United Kingdom, 2007.
[19] B.R. Glasberg, B.C. Moore, Derivation of auditory filter shapes from notched-noise data, Hear. Res. 47 (1–2) (1990) 103–138.
[20] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.
[21] S. Young, et al., The HTK book (for HTK version 3.4), Cambridge University Engineering Department, 2009.
[22] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (86) (2008) 2579–2605.
[23] K.S. Alqudaihi, et al., Cough sound detection and diagnosis using artificial intelligence techniques: challenges and opportunities, IEEE Access 9 (2021) 102327–102344.
[24] S. Khomsay, R. Vanijjirattikhan, J. Suwatthikul, Cough detection using PCA and Deep Learning, in: Proceedings of the International Conference on Information and Communication Technology Convergence (ICTC), Jeju, South Korea, 2019, pp. 101–106.
[25] A. Hassan, I. Shahin, M.B. Alsabek, COVID-19 detection system using recurrent neural networks, in: International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI), United Arab Emirates, Sharjah, 2020, pp. 1–5.
[26] C.R. Finley, et al., What are the most common conditions in primary care? Can. Fam. Physician 64 (11) (2018) 832–840.
[27] P. Webster, Virtual health care in the era of COVID-19, Lancet 395 (10231) (2020) 1180–1181.