Detecting Pertussis in the Pediatric Population Using Respiratory Sound Events and CNN

Roneel V. Sharan[a,*], Shlomo Berkovsky[a], David Fraile Navarro[a], Hao Xiong[a], Adam Jaffe[b,c]

[a]Australian Institute of Health Innovation, Macquarie University, Sydney, NSW 2109, Australia
[b]School of Women's and Children's Health, Faculty of Medicine, University of New South Wales, Sydney, NSW 2052, Australia
[c]Department of Respiratory Medicine, Sydney Children's Hospital, Sydney, NSW 2031, Australia
*Corresponding author. Email: roneel.sharan@mq.edu.au

Abstract

Background and Objective: Pertussis (whooping cough), a respiratory tract infection, is a significant cause of morbidity and mortality in children. The classic presentation of pertussis includes paroxysmal coughs followed by a high-pitched intake of air that sounds like a whoop. Although these respiratory sounds can be useful in making a diagnosis in clinical practice, the distinction of these sounds by humans can be subjective. This work aims to objectively analyze these respiratory sounds using signal processing and deep learning techniques to detect pertussis in the pediatric population.

Methods: Various time-frequency representations of the respiratory sound signals are formed and used as a direct input to convolutional neural networks, without the need for feature engineering. In particular, we consider the mel-spectrogram, wavelet scalogram, and cochleagram representations which reveal spectral characteristics at different frequencies. The method is evaluated on a dataset of 42 recordings, containing 542 respiratory sound events, from children with pertussis and non-pertussis. We use data augmentation to prevent model overfitting on the relatively small dataset and late fusion to combine the learning from the different time-frequency representations for more informed predictions.

Results: The proposed method achieves an accuracy of 90.48% (AUC=0.9501) in distinguishing pertussis subjects from non-pertussis subjects, outperforming several baseline techniques.

Conclusion: Our results suggest that detecting pertussis using automated respiratory sound analysis is feasible. It could potentially be implemented as a non-invasive screening tool, for example, in smartphones, to increase the diagnostic utility for this disease which may be used by parents/carers in the community.

Keywords: Convolutional neural network, cough sound, late fusion, pertussis, time-frequency image

1. Introduction

Pertussis, commonly known as whooping cough, is a respiratory tract infection caused by *Bordetella pertussis* coccobacillus. It spreads by air droplets and is highly contagious [18]. The number of pertussis cases has decreased since the development of a vaccine. However, neither immunization nor previous infection provide lifelong immunity to the disease [2]. There is a resurgence of pertussis infections which is attributed to waning immunity and bacteria mutation [23, 34]. While pertussis affects all age groups, it is a significant cause of morbidity and mortality in young children [35], especially in developing countries, where access to timely diagnoses may not be available.

Following an incubation period, the typical progression of pertussis is in three distinct stages: catarrhal phase, paroxysmal phase, and convalescent phase [18]. The catarrhal phase characteristics are similar to other upper respiratory tract infections. This is followed by the paroxysmal phase. Cough is one of the symptoms of pertussis and it increases in severity at this stage, developing into a paroxysmal or hacking cough followed by a high-pitched intake of air that sounds like a whoop, hence the name whooping cough [35]. The residual cough can persist for weeks to months in the convalescent phase. In severe cases in infants it can lead to respiratory failure and death [20].

People with pertussis are infectious for weeks but, if given appropriate antibiotic treatment, the infectious period and spread is reduced and may also prevent complications [4]. Early treatment of pertussis is, therefore, crucial for managing this disease. We posit that the paroxysmal coughing and whooping sounds can be useful for screening pertussis, especially in the pediatric population which remains the most vulnerable age group. However, recognizing these respiratory sounds by parents/carers of the child can be unfeasible. In clinical practice, this is dependent on the skills and training of the clinicians.

In this work, we aim to develop an objective computational method for detecting respiratory sound events associated with pertussis, that is, the hacking cough and whooping, for the pediatric population. If disseminated widely, for example, as a smartphone application, such an objective assessment tool could prove useful as a screening tool for parents/carers. It could also be useful in developing countries and remote communities which lack access to health facilities and clinicians.

1.1 Related Work

Detecting respiratory diseases using digital respiratory sounds, cough sounds, in particular, has generated interest recently such as in detecting childhood pneumonia [16], monitoring chronic obstructive pulmonary disease [9], and in detecting croup, which is common in children between the age

of 6 months to 6 years and produces a distinctive barking cough [30]. Various signal processing and machine learning techniques have been proposed for the analysis and detection of cough sounds. Being a relatively new area of research, a number of techniques are inspired by other audio classification tasks such as speech recognition. One such measure is mel-frequency cepstral coefficients (MFCCs) [8]. MFCCs utilize mel-filters which are effective in revealing the perceptually significant characteristics of the speech spectrum in small time windows. Speech and cough share some similarities in the generation process and the physiology which could explain the widespread use and effectiveness of MFCCs in cough sound analysis tasks [10, 16, 27, 29, 30, 37].

It is a common practice to complement MFCCs with other techniques. In [10, 16, 29], various temporal and spectral analysis techniques are employed for this purpose. In addition, wavelet transformation is applied in [16] in analysis of cough sounds for detecting pneumonia. Wavelets are effective at the decomposition of non-stationary signals in both the time and frequency domains and, in [16], the focus is particularly on picking the crackle sounds in pneumonia coughs.

Furthermore, the spectral information contained in cough sounds is more dominant in low frequencies than in high frequencies. The human auditory model offers a higher resolution for low frequencies than for high frequencies. In [30], this frequency selectivity property of the human cochlea is modeled using a gammatone filter to differentiate the barking cough sound of croup subjects from the cough sound of other respiratory diseases. A similar approach is also taken in [37].

Audio sound analysis, including cough sound analysis, is typically carried out in small time windows at different frequency localizations. These result in a high dimensional data which conventional classification methods may be unable to handle. A common approach is to reduce this data size into a smaller feature set using statistical methods. With MFCCs, for example, the mean and standard deviation of the coefficients have been used [30]. Similarly, the slope of the wavelet coefficients is used as wavelet feature (WF) in [16]. In [30], the time-frequency representation is formed using gammatone filters, referred to as gammatone-spectrogram or cochleagram, is divided into blocks and the second and third central moments are used as the cochleagram image features (CIF). In [16, 29, 30], feature extraction follows feature selection to further reduce the feature dimension and select the most dominant features for classification.

The use of conventional feature engineering techniques inevitably leads to loss of some information which causes poor classification performance and misdetection of respiratory diseases. More recently, these methods have been superseded by deep learning techniques due to their superior classification results. One such deep learning technique is convolutional neural network (CNN) [17]. CNN is originally an image classification technique which has the ability to learn distinguishing image characteristics directly from the raw image through various mathematical operations. In audio signal classification tasks, this arrangement is typically realized by transforming the signal into an image-like representation [21, 32]. Time-frequency representation of audio signals is the most common approach for this purpose, such as the conventional spectrogram representation formed using short-time Fourier transform (STFT).

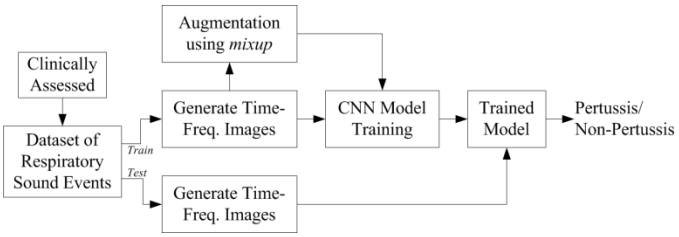

Fig. 1: An overview of the proposed approach in detecting pertussis using respiratory sound events and CNN.

1.2 Audio Data and CNN for Pertussis Detection

An overview of the proposed approach is given in Fig. 1. We take inspiration from conventional feature extraction techniques and the state of the art CNN for detecting pertussis using respiratory sounds. In particular, we represent the one-dimensional respiratory sound signals as two-dimensional time-frequency representations for classification using CNN. Our approach in forming the time-frequency representations is based on the feature extraction techniques from [16, 29, 30]. In particular, we use mel-filters, as used in computing MFCCs, to form *mel-spectrogram*; wavelet transform, as used in computing WF, to form wavelet *scalogram*, and gammatone filters, as used in computing CIF, to form *cochleagram*.

Furthermore, different time-frequency representations reveal spectral characteristics at different frequencies. In conventional machine learning, this information is combined, for example, using feature vector concatenation, to improve the classification performance. With CNN this can be achieved using late fusion whereby the outputs of CNN models trained on different representations are combined. This can be realized either by averaging the output scores [39] or using the output scores to train a secondary classifier [41]. In this work, we use late fusion to combine the CNN learning from different time-frequency representations, aiming to make more accurate predictions.

The proposed approach is evaluated on a dataset of respiratory sounds from children with suspected or confirmed pertussis and other respiratory diseases. Collecting physiological data is time consuming, expensive, and may require patient cooperation, which can be difficult with children. However, a rapid rise in the use of digital technology has prompted researchers to collect self-reported data from the public. In a similar study [29], researchers composed a dataset of respiratory diseases using online sources while researchers at Microsoft used web search queries of users with self-identified conditions [36]. More recently, researchers at the University of Cambridge collected COVID-19 related sounds of users with self-reported disease status through a website and a smartphone application. In this work, we use a dataset of respiratory sounds collated from the YouTube online video sharing platform and reviewed by a clinician.

Table 1: An overview of the dataset used in this work.

Disease Group	Number of Recordings	Total Recording Duration (sec)	Age Known	Age Range (Mean) (Months)	Gender Known (M:F)	Total Number of Coughs (Range)	Subjects with Whoops (Total Whoops)
Pertussis	21	1261.23	15	1–84 (25.27±24.35)	18 (8:10)	545 (3–100)	16 (110)
Asthma	2	1674.62	7	5–36 (18.36±14.18)	16 (11:5)	257 (2-47)	0 (0)
Bronchiolitis	6						
Croup	12						
Pneumonia	1						

In total, the dataset contains 42 recordings, each with multiple respiratory sounds. This makes it a relatively small dataset and CNN models trained on small datasets can be prone to overfitting. One method to reduce overfitting is *mixup* [40] which augments the dataset, mixing the features of different classes. It is a simple yet effective method with very low computational costs. In this work, we extend the mixup data augmentation technique to time-frequency representations of respiratory sounds.

The rest of the paper is organized as follows. An overview of the dataset and the proposed method is given in Section 2. The experimental setup and results are provided in Section 3 and discussion of the results and conclusions are in Section 4.

2 Methods

2.1 Dataset

The dataset used in this work was collated from YouTube. Various search terms were used to identify respiratory sound recordings from children with the following respiratory conditions: pertussis, asthma, bronchiolitis, croup, and pneumonia. The diagnosis of pertussis and other respiratory conditions in the videos was attributed by the information provided in the title and/or description of the videos and later checked by a clinician to assess the plausibility of the sounds and the reported diagnosis.

The final dataset contained a total of 42 recordings, each with multiple respiratory sound events. The breakdown of disease classes for the recordings in the dataset is as follows: 21 pertussis, 2 asthma, 6 bronchiolitis, 12 croup, and 1 pneumonia (in total, 21 non-pertussis). The age and gender of the subjects were determined based on the information provided in the title and description of the videos and, if needed, by watching the video. While all the subjects were children, the age and gender could not be determined in some cases. A summary of the respiratory diseases, demographics, and breakdown of respiratory sound events in the dataset is provided in Table 1.

All subjects had cough as a symptom with the average number of coughs in the pertussis group more than twice in the non-pertussis group despite the average recording duration for pertussis subjects being lower than for non-pertussis subjects. This could be attributed to the nature of the paroxysmal cough in pertussis – persistent hacking cough followed by a whoop. However, whooping is not always present in pertussis subjects. In this dataset, 16 of the 21 pertussis subjects were determined to have whoops with a total of 110 whooping episodes.

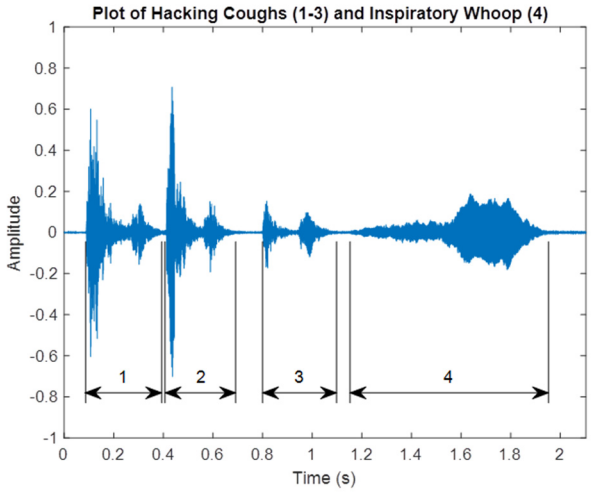

Fig. 2: Plot of three hacking cough signals followed by an inspiratory whoop from a pertussis subject.

An illustration of hacking cough signals followed by a whoop from a pertussis subject is given in Fig. 2. In this case, the subject has three rapid bursts of cough over a period of about 1 second, marked as 1-3. Looking at the amplitude of the signals, we observe that bulk of the air is expelled from the lungs of the subject in the first two coughs. The subject may be almost out of air at the time of the third cough which explains the significantly lower amplitude of the cough signal. There is an urgent need to breathe after that with the subject having a long gasp for air sounding like a whoop, marked as 4.

Most, if not all, the recordings are believed to be captured using smartphone microphones. This means the recordings are likely coming from different manufacturers and models of smartphones making this a challenging and diverse data. Of the 42 recordings, 23 are recorded in a home environment, 2 in hospital, 2 in vehicle, 5 in an unknown indoor environment while the location of 10 recordings is unknown. The recordings include various background noises and sound events such as people talking, TV or music playing, dog barking, noise from other household appliances, etc. The signal-to-noise ratio (SNR) in the presence of background noise is also diverse, estimated to be in the range of 16 dB to 44 dB.

The audio data were retrieved in the MP3 coding format. The sampling frequency of the original recording is not known but the retrieved audio files are sampled at 44.1 kHz. Inspecting the spectrogram of the audio files revealed that some recordings did not have any spectral information above 8 kHz. Hence, spectral analysis was limited to this frequency.

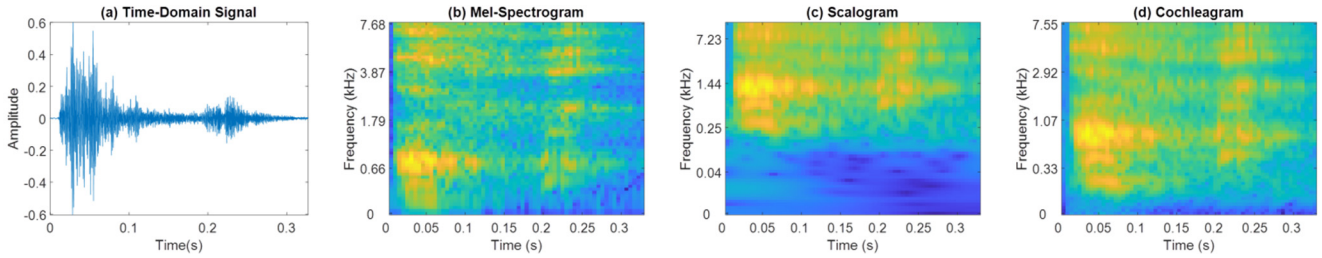

Fig. 3: Plot of (a) cough sound signal, (b) mel-spectrogram representation, (c) wavelet scalogram representation, and (d) cochleagram representation of this cough signal.

The noise source generally varies from one signal to another and most noise types were present in the targeted 0-8 kHz frequency range, therefore, no specific noise filtering was performed.

2.2 Time-Frequency Representations

In the proposed method, we use time-frequency representations of respiratory signals as a direct input to the CNN. Three time-frequency representations are considered for this purpose: mel-spectrogram, wavelet scalogram, and cochleagram. In this work, a target time-frequency image size of 64×64 is used.

In forming the *mel-spectrogram*, STFT is computed by dividing the respiratory signal into 64 frames with a 50% overlap and computing the Fourier transform as

$$X_F(k) = \sum_{n=0}^{N-1} x(n) e^{-\frac{i2\pi}{N}kn} \tag{1}$$

where N is the length of the frame, $x(n)$ is the framed respiratory signal, and $X(k)$ is the k^{th} harmonic frequency. The log mel-filter bank energies are then computed as

$$E(m) = \log \sum_{k=0}^{\frac{N}{2}-1} V(m)\left|X(k)\right|, \qquad m = 1,2,\ldots,M \tag{2}$$

where $V(m)$ is the normalized filter response of the mel-filters, triangular filter banks equally spaced on the mel-scale [26]. The number of mel-filters M is set to 64 and this process is repeated for each of the 64 frames resulting in a 64×64 mel-spectrogram.

To form the *wavelet scalogram*, continuous wavelet transform is applied to the respiratory signal $x(t)$ at scale $e > 0$ and position h as

$$X_W(e,h) = \frac{1}{\sqrt{e}} \int_{-\infty}^{\infty} x(t) \bar{\psi}\left(\frac{t-h}{e}\right) dt \tag{3}$$

where ψ is the mother wavelet (complex conjugate) [19]. The scale parameter is analogous to frequency in Fourier transform and is chosen such that the signal is decomposed into 64 components. The decomposed signal at each scale is divided into 64 windows, the absolute values of which are summed to determine the energy in each window and log of the energy values gives the wavelet scalogram.

The frequency components of the *cochleagram* are based on the human auditory filters, modeled by a gammatone filter, the impulse response of which is given as

$$g(t) = at^{j-1} e^{-2\pi bt} \cos\left(2\pi f_c t + \phi\right) \tag{4}$$

where a is the amplitude, t is the time, j is the filter order, f_c is the center frequency of the filter, b is the bandwidth of the filter, and ϕ is the phase factor [28]. The relationship between the center frequency and the bandwidth of the filters in human hearing is described using the equivalent rectangular bandwidth (ERB) [28]. The implementation of the ERB filter model given in [11] was exploited. In forming the cochleagram, the respiratory signal was filtered using 64 gammatone filters, followed by the windowing process similar to wavelet scalogram.

Illustrations of a cough signal and its mel-spectrogram, scalogram, and cochleagram representations are given in Fig. 3 with a frequency range of 0-8 kHz.

2.3 CNN

2.3.1 CNN Architecture

The architecture of the 2-D CNN is similar to the one presented in [31]. The input layer has a size of 64×64. The time-frequency representations are normalized using zero mean and unit standard deviation in the input layer to remove the effect of subject variability. The classification network consists of five convolutional layers, each with a 3×3 filter size. The number of filters in the first convolutional layer is $N_F = 48$, with $2N_F$ filters in the second layer, and $4N_F$ filters in the remaining three layers.

Each convolutional layer is followed by a batch normalization layer [12] and rectified linear unit (ReLU) [24]. These are followed by a max pooling layer [14] in all the layers, except for the fourth layer. Each max pooling layer has a pool size of 3×3 and a stride of 2×2. The final max pooling layer is followed by a dropout layer [33] with probability 0.2, a fully connected layer, and a softmax layer [3].

The number of pertussis labels in the training dataset is greater than that of non-pertussis labels. To account for this imbalance, weighted cross entropy loss was used in the

classification layer. Given the prediction scores Y and training targets T, the weighted cross entropy loss is computed as

$$L = -\frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{K} c_i T_{si} \log\left(Y_{si}\right) \tag{5}$$

where S is the number of observations, K is the number of classes, and c are the class weights. The final network, therefore, has a total of 24 layers.

The network was trained using adaptive moment estimation (Adam) [15] which uses the estimates of the first and second moments of the gradients to compute adaptive learning rate for the parameters. The first and second moment estimators for training iteration t are computed as

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t} \text{ and } \hat{v}_t = \frac{v_t}{1-\beta_2^t}, \tag{6}$$

respectively, where β_1 and β_2 are the hyperparameters of the algorithm. The model weight w is then updated as

$$w_t = w_{t-1} - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \varepsilon} \tag{7}$$

where η is the step size and ε is a small scalar.

The training parameters are set using a simple grid search. The initial learn rate is set to 0.003, mini batch size is 32, and the maximum number of epochs is 30. In addition, we use a learn rate drop factor of 0.1, learn rate drop period of 10, and L_2 regularization of 0.2. The model was implemented in MATLAB R2020b and trained on AWS using a single NVIDIA V100 Tensor Core GPU. The training stops after the maximum number of epochs is reached.

2.3.2 Data Augmentation

As our training dataset is relatively small, we perform data augmentation using a modified version of mixup to prevent overfitting. Given a training dataset of time-frequency representations D, for every representation I_r belonging to class y_r, additional representation $\tilde{I}_r$ is created by mixing I_r with a randomly selected representation I_s from another class y_s, as

$$\tilde{I}_r = \lambda I_r + (1-\lambda) I_s, \quad \tilde{y}_r = y_r \tag{8}$$

where, $\lambda = 0.5$, $I_r \neq I_s$, and $y_r \neq y_s$. We used this process to create a mixup time-frequency representation for each time-frequency image in D. The original dataset D and the mixup dataset $\tilde{D}$ are combined to train the CNN model.

2.3.3 Late Fusion

Furthermore, the three time-frequency representations reveal spectral characteristics at slightly different center frequencies and bandwidths; refer to Fig. 3 (b)-(d). As such, a CNN trained on these time-frequency representations would learn unique information and combining the learnings from the CNNs has the potential to improve the classification performance. In this work, we use a late fusion approach for this purpose, as illustrated in Fig. 4.

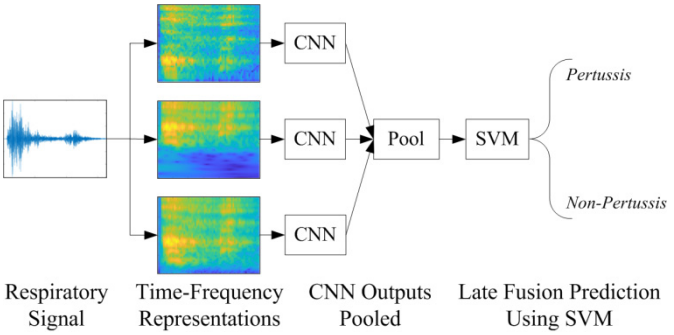

Fig. 4: Late fusion by pooling the CNN outputs for a combined prediction using SVM.

Each of the three CNNs, trained on mel-spectrogram, wavelet scalogram, and cochleagram, outputs a probability score p_1, p_2, and p_3, respectively, for each validation sample. These probability scores are concatenated into a feature vector $[p_1, p_2, p_3]$. These feature vectors are used to train a secondary classifier, in this case a support vector machine (SVM) [6] with radial basis function (RBF) kernel. The same training and validation procedure, as with CNN, is repeated to make the final prediction for each validation sample.

2.4 Evaluation Metrics

Performance of the methods is evaluated using sensitivity (Sen), specificity (Spe), accuracy (Acc), and the area under the curve (AUC) of the receiver operating characteristic (ROC) curve given as follows:

$$Sen = \frac{TP}{TP+FN}, \; Spe = \frac{TN}{TN+FP}, \tag{9}$$

$$Acc = \frac{TP+TN}{TP+FN+TN+FP}, \text{ and} \tag{10}$$

$$AUC = \int_0^1 f(x)\,dx, \tag{11}$$

where TP, FP, TN, and FN are the number of true positives, false positives, true negatives, and false negatives, respectively, and $f(x)$ is the ROC function curve. We computed the AUC using trapezoidal approximation.

In similar to [30], the performance is evaluated at the cough and subject levels. At the cough level, the aim is to correctly predict the class of each cough as pertussis or non-pertussis using the posterior probability of the classification models. The u^{th} cough from the v^{th} subject/recording is classified as pertussis if the posterior probability is greater than or equal to t_1, that is,

Table 2: Results using baseline methods.

Feature Set	Classifier	Cough Classification Results				Subject Classification Results			
		Sen (%)	Spe (%)	Acc (%)	AUC	Sen (%)	Spe (%)	Acc (%)	AUC
MFCC	LR	63.85	63.82	63.84	0.6659	85.71	**80.95**	83.33	0.7880
	NB	60.06	59.80	59.96	0.6422	71.43	66.67	69.05	0.6859
	SVM	**69.97**	**69.85**	**69.93**	**0.7242**	85.71	**80.95**	83.33	**0.8787**
WF	LR	56.85	56.78	56.83	0.5977	80.95	71.43	76.19	0.7800
	NB	65.89	63.32	64.94	0.6585	85.71	71.43	78.57	0.7029
	SVM	62.39	62.31	62.36	0.6525	85.71	**80.95**	83.33	0.8435
CIF	LR	60.06	59.80	59.96	0.6288	80.95	71.43	76.19	0.7789
	NB	60.93	60.80	60.89	0.6428	76.19	71.43	73.81	0.7721
	SVM	64.43	64.32	64.39	0.6588	85.71	**80.95**	83.33	0.8129
MFCC + WT + CIF	LR	60.06	59.80	59.96	0.6340	85.71	76.19	80.95	0.8061
	NB	63.56	63.32	63.47	0.6829	76.19	71.43	73.81	0.7120
	SVM	62.39	62.31	62.36	0.6703	**90.48**	**80.95**	**85.71**	0.8141

$$C_v^u = \begin{cases} 1, & p \geq t_1 \\ 0, & p < t_1 \end{cases} \quad (12)$$

where $u \in \mathbb{Z}: u \in [1, U_v]$, U_v is the total number of coughs for the v^{th} subject, $v \in \mathbb{Z}: v \in [1, V]$, and V is the total number of subjects. The optimal cut-off point on the ROC curve t_1 is chosen at the intersection of cough level sensitivity and specificity.

At the subject level, the classification of the coughs is used to determine if the recording belongs to a subject with pertussis or non-pertussis as

$$R_v = \begin{cases} 1, & q_v \geq t_2 \\ 0, & q_v < t_2 \end{cases} \quad (13)$$

where $q_v = {}^1/_{U_v} \sum_{u=1}^{U_v} C_v^u$ and the optimal cut-off point t_2 is chosen at the intersection of subject level sensitivity and specificity.

3 Experimental Evaluation

3.1 Experimental Setup

In this work, we use a stratified 7-fold cross-validation which we found to give a good compromise between the number of training and validation samples in each fold. As such, 3 pertussis and 3 non-pertussis recordings are used for validating the model and the remaining recordings are used for training the model, in each fold. The respiratory sounds from a recording/subject are present either in the training or validation dataset, but not in both.

The number of respiratory sounds per recording varies from 2 to 138. When using all the available respiratory sounds, the model naturally tends to fit to the recordings with more respiratory sounds. This means that even though the cough classification results improve, this does not necessarily translate to a better subject classification as the model may not be adequately learning the inter-subject variability. To address this, we varied the maximum number of respiratory sounds per subject to be used in training with a value of 20 yielding the best performance in subject classification. As such, the dataset used in all the experiments contains 343 respiratory sound events from pertussis subjects and 199 events from non-pertussis subjects, a total of 542 events.

We present the evaluation results in Section 3.2 and 3.3.

3.2 Baseline Method and Results

The baseline features experimented with in this work are mel-frequency cepstral coefficients (MFCC), wavelet features (WF), and cochleagram image features (CIF), and their combined feature set. To compute MFCCs, the respiratory signal is divided into frames of 1024 points with 50% overlap. After applying Fourier transform, 20 mel-filter bank energies are computed in each frame followed by discrete cosine transform [13] to obtain the MFCCs. The first and second derivatives of the coefficients are also computed [38]. The final 120 dimensional MFCC feature vector is the mean and standard deviation of the 20 coefficients and the derivatives across all frames.

To compute WF [16], a time-scale representation or scalogram is formed, in similar to the procedure described in Section 2.2 but with an output size of 64×12. The final 768 dimensional feature vector is the slope between the energy values in the 12 windows in each time axis. The cochleagram representation described in Section 2.2 is used to compute the CIF [30]. The cochleagram is divided into 8 blocks along the time and frequency axis, and second and third central moments are computed in each of the 64 blocks to form a 128 dimensional feature vector.

Following feature extraction, the significance of the features is determined using training data in each fold with one-way analysis of variance (ANOVA) and *t*-test. One-way ANOVA and *t*-test determine if there is a significant difference between the mean of the two groups, pertussis and non-pertussis. The significance of each feature dimension is given by the *p*-value in the range [0,1] where a *p*-value close to 0 indicates high significance and a *p*-value close to 1 indicates low significance. In similar to [1, 30], various *p*-value thresholds are applied in the range [0,1] and feature dimensions with

Table 3: Results using the proposed method.

Input	Classifier	Cough Classification Results				Subject Classification Results			
		Sen (%)	Spe (%)	Acc (%)	AUC	Sen (%)	Spe (%)	Acc (%)	AUC
Mel-Spectrogram	CNN	72.30	71.86	72.14	**0.7962**	90.48	80.95	85.71	0.9172
Scalogram	CNN	72.01	71.86	71.96	0.7767	85.71	80.95	83.33	0.8741
Cochleagram	CNN	71.14	70.85	71.03	0.7705	90.48	80.95	85.71	0.8730
Late Fusion	SVM	**73.18**	**72.86**	**73.06**	0.7640	**95.24**	**85.71**	**90.48**	**0.9501**

p-value below this threshold are used for training and validation. The baseline classifiers used in this work are logistic regression (LR) [7], Naïve Bayes (NB) [22], and SVM with RBF kernel, as seen to be popular in cough sound classification tasks [16, 30, 37].

Results for pertussis and non-pertussis classification at cough and subject level using the baseline methods are given in Table 2. While we experimented with both the feature selection methods and several *p*-value thresholds, only the best results are presented here. In general, the best baseline results for the cough and subject classifications are achieved by SVM. At the cough level, the best accuracy for a single feature set is in the range of 64.39% to 69.93% and the best AUC is in the range of 0.6585 to 0.7242. With a cough level accuracy of 69.93% (AUC=0.7242) and a subject level accuracy of 83.33% (AUC=0.8787), the best classification performance using single feature set is with MFCC and SVM. The combined feature set yields mixed performance. At the cough level, it could not improve on the MFCC-SVM results. The accuracy improves at subject level to 85.71% but with a lower AUC of 0.8141.

3.3 Results Using Proposed Method

The classification results for time-frequency image classification using CNN and late fusion are given in Table 3. Target time-frequency image size of 64×64 is used as higher dimensional images increased computational overheads but did not improve the classification results. In late fusion, the output of the individual CNNs trained on the three time-frequency representations is combined for classification using SVM and evaluated in 7-fold cross validation.

All three time-frequency representation classification using CNN achieve accuracy greater than 71% and AUC greater than 0.77 in cough classification. This is substantially higher than the results using a single baseline feature set or the combined MFCC+WF+CIF. With a sensitivity of 72.30%, specificity of 71.86%, accuracy of 72.14%, and AUC of 0.7962, the best results in cough classification are achieved using mel-spectrogram-CNN, while the accuracy and AUC scores of cochleagrams and scalograms are close. Mel-spectrograms also produce the best subject level classification results: sensitivity of 90.48%, specificity of 80.95%, accuracy of 85.71%, and AUC of 0.9172. The combination of cochleagram and CNN also achieves an accuracy of 85.71% for subject level classification but with a lower AUC.

With late fusion, the cough level sensitivity, specificity, and accuracy improve to 73.18%, 72.86%, and 73.06%, respectively, but with a lower AUC of 0.7640. In addition, the subject classification results improve with a sensitivity of 95.24%, specificity of 85.71%, accuracy of 90.48%, and AUC of 0.9501. These are the best overall results in detecting pertussis and non-pertussis subjects. Using the Wilson method [25], the 95% confidence interval for sensitivity is 77.33% to 99.15% and for specificity 71.09% to 97.35%. In addition, the classification accuracy is 90% at SNR below 25 dB, 86.96% at SNR of 25 dB to 35 dB, and 100% over 35 dB which indicates better classification accuracy when the noise level is low, as can be expected.

4 Discussion and Conclusions

The dataset used in this work has been recorded in natural environments with SNR as low as 16 dB. The recordings are believed to be made using smartphones of different manufacturers and models and the training and validation procedure followed in this work is subject independent. All these increase the difficulty and complexity of the task. Despite these constraints, our method is empirically shown to achieve strong classification performance at the cough and particularly subject levels. In addition, while earlier works looked at the problem of respiratory sound-based pertussis detection using conventional feature engineering and machine learning methods [27, 29], our proposed time-frequency representations, CNN, and late fusion approach forgoes the need for feature engineering and outperforms several conventional methods. These demonstrate the robustness of the proposed approach against the diversity of recording environments, background noises, and recording devices, and against conventional classification methods.

Our work has some limitations. Further analysis of our results shows that 3 out of the 4 misclassifications for the best classification model are in children aged 6 months or less. This could be because the lungs and airway muscles of children are in different developmental stages at different age groups. This may cause variations in the cough sound, especially in infancy [5]. Age group specific models may help overcome this problem, however, how exactly the sound is affected by age or the adequacy of specific cut-off point to establish age groups remain unclear. In addition, in this work we have only 42 subjects of which the age of only 22 subjects is known. The relatively small dataset makes this difficult. Furthermore, the non-pertussis group in the dataset is comprised of a number of different respiratory diseases of which pneumonia and asthma have only 1 and 2 recordings, respectively. Also, the dataset does not include other types of childhood respiratory diseases or comorbidities which would be present in complex cases. While our dataset is still larger

than those of [27, 29], the availability of an even larger and complex data with age-defined groups would help us develop more generalizable models. Moreover, *Bordetella parapertussis* causes a similar clinical picture to *Bordetella pertussis* but tends to be milder and of a shorter duration. However, we were not able to verify the diagnosis of pertussis using confirmed microbiology and relied on physician interpretation as the gold standard. We hope to collect clinically verified data in future prospective studies.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declarations of interest: none.

References

[1] U.R. Abeyratne, V. Swarnkar, A. Setyati, R. Triasih, Cough sound analysis can rapidly diagnose childhood pneumonia, Annals of biomedical engineering, 41 (11) (2013) 2448-2462.

[2] American Academy of Pediatrics, Pertussis (Whooping Cough), in: L.K. Pickering, C.J. Baker, D.W. Kimberlin, S.S. Long (Eds.) Red Book: 2012 Report of the Committee on Infectious Diseases, American Academy of Pediatrics, Elk Grove Village, IL, 2012.

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

[4] Centers for Disease Control and Prevention, Pertussis (Whooping Cough), 2019. [Online]. Available: www.cdc.gov [Accessed: 2nd September, 2020]

[5] A.B. Chang, J.G. Widdicombe, Cough throughout life: Children, adults and the senile, Pulmonary Pharmacology & Therapeutics, 20 (4) (2007) 371-382.

[6] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning, 20 (3) (1995) 273-297.

[7] J.S. Cramer, The origins of logistic regression, Tinbergen Institute, Discussion Paper 2002-119/4, 2002.

[8] S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Transactions on Acoustics, Speech and Signal Processing, 28 (4) (1980) 357-366.

[9] A.C. den Brinker, R. van Dinther, M.G. Crooks, S. Thackray-Nocera, A.H. Morice, Alert system design based on experimental findings from long-term unobtrusive monitoring in COPD, Biomedical Signal Processing and Control, 63 (2021) 102205.

[10] T. Drugman, Using mutual information in supervised temporal event detection: Application to cough detection, Biomedical Signal Processing and Control, 10 (2014) 50-57.

[11] D.D. Greenwood, A cochlear frequency-position function for several species - 29 years later, Journal of the Acoustical Society of America 87 (6) (1990) 2592-2605.

[12] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167, (2015).

[13] A.K. Jain, Fundamentals of Digital Image Processing, Prentice Hall, Englewood Cliffs, NJ, 1989.

[14] K. Jarrett, K. Kavukcuoglu, M.A. Ranzato, Y. LeCun, What is the best multi-stage architecture for object recognition?, in: IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 2009, pp. 2146-2153.

[15] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, (2014).

[16] K. Kosasih, U.R. Abeyratne, V. Swarnkar, R. Triasih, Wavelet augmented cough analysis for rapid childhood pneumonia diagnosis, IEEE Transactions on Biomedical Engineering, 62 (4) (2015) 1185-1194.

[17] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems (NIPS), 2012, pp. 1097-1105.

[18] A.M. Lauria, C.P. Zabbo, Pertussis (whooping cough), StatPearls Publishing, Treasure Island (FL), 2020.

[19] S. Mallat, A Wavelet Tour of Signal Processing: The Sparse Way, Academic Press, United States, 2009.

[20] J.A. Melvin, E.V. Scheller, J.F. Miller, P.A. Cotter, Bordetella pertussis pathogenesis: Current and future challenges, Nature Reviews Microbiology, 12 (4) (2014) 274-288.

[21] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, M.D. Plumbley, Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26 (2) (2018) 379-393.

[22] M. Minsky, Steps toward artificial intelligence, Proceedings of the IRE, 49 (1) (1961) 8-30.

[23] F.R. Mooi, N.A.T. Van Der Maas, H.E. De Melker, Pertussis resurgence: Waning immunity and pathogen adaptation – two sides of the same coin, Epidemiology and Infection, 142 (4) (2014) 685-694.

[24] V. Nair, G.E. Hinton, Rectified linear units improve restricted boltzmann machines, in: 27th International Conference on Machine Learning, Haifa, Israel, 2010, pp. 807-814.

[25] R.G. Newcombe, Two-sided confidence intervals for the single proportion: Comparison of seven methods, Statistics in Medicine, 17 (8) (1998) 857-872.

[26] D. O'Shaughnessy, Speech Communication: Human and Machine, Addison-Wesley Pub. Co., 1987.

[27] D. Parker, J. Picone, A. Harati, S. Lu, M.H. Jenkyns, P.M. Polgreen, Detecting paroxysmal coughing from pertussis cases using voice recognition technology, PLOS ONE, 8 (12) (2013) e82971.

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

[29] R.X.A. Pramono, S.A. Imtiaz, E. Rodriguez-Villegas, A cough-based algorithm for automatic diagnosis of pertussis, PLOS ONE, 11 (9) (2016) 1-20.

[30] R.V. Sharan, U.R. Abeyratne, V.R. Swarnkar, P. Porter, Automatic croup diagnosis using cough sound recognition, IEEE Transactions on Biomedical Engineering, 66 (2) (2019) 485-495.

[31] R.V. Sharan, S. Berkovsky, S. Liu, Voice command recognition using biologically inspired time-frequency representation and convolutional neural networks, in: 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Montreal, QC, Canada, 2020, pp. 998-1001.

[32] R.V. Sharan, T.J. Moir, Acoustic event recognition using cochleagram image and convolutional neural networks, Applied Acoustics, 148 (2019) 62-66.

[33] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, Journal of Machine Learning Research, 15 (2014) 1929-1958.

[34] A.A.J. van der Ark, D.F. Hozbor, C.J.P. Boog, B. Metz, G.P.J.M. van den Dobbelsteen, C.A.C.M. van Els, Resurgence of pertussis calls for re-evaluation of pertussis animal models, Expert Review of Vaccines, 11 (9) (2012) 1121-1137.

[35] World Health Organization, Pertussis vaccines: WHO position paper – August 2015, 2015.
[36] E. Yom-Tov, Screening for cancer using a learning internet advertising system, ACM Transactions on Computing for Healthcare, 1 (2) (2020) Article 10.
[37] M. You, Z. Liu, C. Chen, J. Liu, X.-H. Xu, Z.-M. Qiu, Cough detection by ensembling multiple frequency subband features, Biomedical Signal Processing and Control, 33 (2017) 132-140.
[38] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, The HTK book (for HTK version 3.4), Cambridge University Engineering Department, 2009.
[39] H. Zeinali, L. Burget, J. Cernocky, Convolutional neural networks and x-vector embedding for DCASE2018 acoustic scene classification challenge, in: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018), Surrey, UK, 2018, pp. 202-206.
[40] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, in: International Conference on Learning Representations, Vancouver, Canada, 2018, pp. 1-13.
[41] W. Zheng, J. Yi, X. Xing, X. Liu, S. Peng, Acoustic scene classification using deep convolutional neural network and multiple spectrograms fusion, in: Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017), Munich, Germany, 2017, pp. 133-137.