Detecting Cough Recordings in Crowdsourced Data Using CNN-RNN

Roneel V. Sharan
Australian Institute of Health Innovation
Macquarie University
Sydney, Australia
roneel.sharan@mq.edu.au

Hao Xiong
Australian Institute of Health Innovation
Macquarie University
Sydney, Australia
hao.xiong@mq.edu.au

Shlomo Berkovsky
Australian Institute of Health Innovation
Macquarie University
Sydney, Australia
shlomo.berkovsky@mq.edu.au

***Abstract*—The sound of cough is an important indicator of the condition of the respiratory system. Automatic cough sound evaluation can aid the diagnosis of respiratory diseases. Large crowdsourced cough sound datasets have recently been used by several groups around the world to develop cough classification models. However, not all recordings in these datasets contain cough sounds. As such, it is important to screen the recordings for the presence of cough sounds before developing cough classification models. This work proposes a method to screen crowdsourced audio recordings for cough sounds using deep learning methods. The proposed approach divides the audio recording into overlapping frames and converts each frame into a mel-spectrogram representation. A pretrained convolutional neural network for audio classification is trained to learn the spectral characteristics of cough and non-cough frames from its mel-spectrogram representation. It is combined with a recurrent neural network to learn the dependencies between the sequence of frames. The proposed method is evaluated on 400 crowdsourced audio recordings, manually annotated as cough or non-cough. An accuracy of 0.9800 (AUC of 0.9973) is achieved in classifying cough and non-cough recordings using the proposed method. The trained network is used to analyze the remaining audio recordings in the dataset, identifying only about 67% of recordings as containing usable cough sounds. This shows the need to exercise caution when using crowdsourced cough data.**

Keywords—cough sound, crowdsourced, deep learning, mel-spectrogram, respiratory diseases

I. Introduction

Cough is a common symptom of respiratory diseases and the sound of cough, such as productive (wet) and non-productive (dry) cough [1], the barking cough of croup [2], and the hacking coughs and inspiratory whoops in pertussis [3], are important predictors of respiratory diseases. Cough sound interpretation in clinical practice can be subjective, depending on the training and skills of the clinician. Hence, objective cough sound evaluation using signal processing and machine learning techniques has the potential to aid the clinician in respiratory disease diagnosis.

Several groups around the world have studied the cough sound of COVID-19, as a possible screening tool for this respiratory disease [4-6]. Collating clinically verified cough data for such a purpose can be time consuming and expensive. These studies have, therefore, relied on crowdsourced data. For cough sound analysis in [2], the cough sound signals are automatically segmented before feature extraction and classification. This way, recordings with no cough sounds can be removed from analysis. However, in [4, 7], the audio recordings are directly input into the feature extraction algorithms. Analysis by clinicians on a subset of recordings from one such large crowdsourced dataset shows that many audio recordings do not contain cough sounds [6]. As such, it is important to automatically screen large crowdsourced datasets of audio recordings for the presence of cough sounds before developing cough sound classification algorithms.

This work proposes a method for detection of cough and non-cough recordings in crowdsourced data using supervised deep learning. Supervised deep learning models can be data hungry and training them from scratch requires a large labeled dataset. Data annotation is, however, time consuming. This work, therefore, makes use of fine-tuning whereby weights of pretrained audio classification networks are updated by retraining on a small manually annotated set of cough and non-cough recordings from a large crowdsourced dataset. This gives faster convergence of the model [8].

Two popular pretrained convolutional neural networks (CNNs), YAMNet and VGGish [9], are studied for this purpose. While these CNNs learn the spectral characteristics within a segment or frame of the audio recordings, they are combined with a bidirectional long short-term memory network (BiLSTM) [10], a type of recurrent neural network (RNN), to learn the dependencies between a sequence of frames in an audio recording. The proposed method outperforms various baseline methods in cough vs. non-cough recording classification and identifies a large proportion of unusable audio recordings amongst other recordings in the dataset. Our work highlights the need to rigorously apply data cleansing and filtering when using crowdsourced data for developing cough sound classification models.

II. Method

A. Dataset

This work utilizes the COUGHVID dataset [6], a crowdsourced data of audio recordings collected for a study of COVID-19 cough sounds. The dataset has 27,550 recordings submitted by subjects using a web application. The recordings were collected all over the world, therefore, representing substantial demographic diversity.

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

978-1-6654-2899-6/21/$31.00 ©2021 IEEE

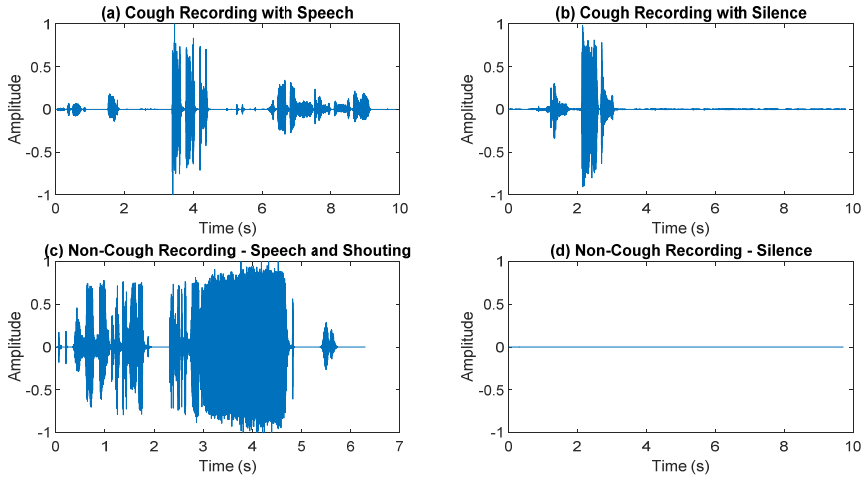

Fig. 1. Illustrative examples of cough and non-cough recordings in the dataset: (a) cough recording with three coughs between 3 and 5 seconds and speech on either side, (b) cough recording with no non-cough sounds, (c) non-cough recording with speech and shouting, and (d) non-cough recording with silence.

TABLE I. OVERVIEW OF THE DATASET USED IN THIS WORK

		Cough	**Non-Cough**	**Overall**
Number of recordings		200	200	400
Number of cough samples		915	0	915
Num. of frames (0.975s, 50% OL)		3485	3535	7020
Gender	Male	123	35	158
	Female	53	38	91
	Other/Unknown	24	127	151
Age (years)		36.24±14.01	39.54±17.64	37.08±15.05

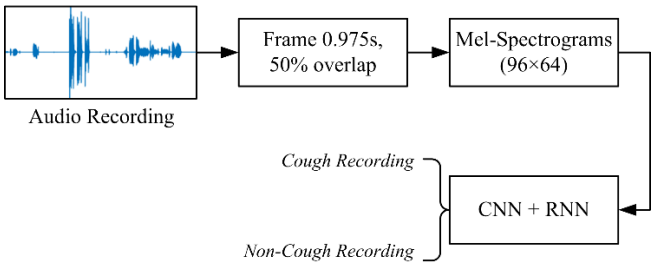

Fig. 2. Overview of the proposed method in cough vs. non-cough recording classification.

In this work, a subset of 400 recordings from the COUGHVID dataset is used to develop cough vs. non-cough recording classification algorithms. These recordings are manually annotated by the first author whereby a recording was labeled as *cough* if it contained one or more cough sounds and *non-cough* otherwise. Illustration of two cough recordings used in this work is given in Fig. 1(a) and Fig. 1(b), and two non-cough recordings in Fig. 1(c) and Fig. 1(d). In Fig. 1(a), there are three cough sounds between 3 and 5 seconds, with speech before and after these coughs. There are no non-cough sounds in the cough recording of Fig. 1(b). The non-cough recording in Fig. 1(c) contains speech and shouting while the non-cough recording in Fig. 1(d) contains only silence. In addition to annotating each recording, the cough sounds in the cough recordings are manually segmented using Audacity (www.audacityteam.org), a digital audio editing software, to determine the start and end point of each cough.

An overview of the final dataset used in this work is provided in Table I. The recordings in the dataset are resampled at 16 kHz. The dataset has a total of 200 cough recordings, containing 915 cough sounds, and 200 non-cough recordings. The age of the subjects, where reported, is 37.08±15.05 years with 158 male and 91 female subjects while 151 subjects did not report their gender.

B. *Proposed Method*

This work proposes a *sequence-to-label* approach for classification of cough vs. non-cough recordings. An overview of the proposed method is illustrated in Fig. 2. The audio recordings are divided into frames of 0.975 seconds with a 50% overlap between adjacent frames. Each frame is transformed into a mel-spectrogram representation [11] of size 96×64. When computing the mel-spectrogram representation, short-time Fourier transformation is performed first which results in a discrete Fourier transform (DFT), $X_m[k]$, for the m^{th} frame. The DFT values are grouped in critical bands of the triangular overlapping filters mapped on the mel-scale [12]. The mel-spectrum of the m^{th} frame for filters $r = 1, 2, \ldots, R$ is defined as

$$\mathrm{MF}_m[r] = \frac{1}{A_r} \sum_{k=L_r}^{U_r} \left| V_r[k] X_m[k] \right|^2, \qquad r = 1, 2, \ldots, R \tag{1}$$

where $V_m[k]$ is the weighting function for the r^{th} filter from DFT index L_r to U_r, and

$$A_r = \sum_{k=L_r}^{U_r} \left| V_r[k] \right|^2 \tag{2}$$

is a normalizing factor for the r^{th} mel-filter. The mel-spectrogram is then computed as the log of the mel-spectrum values in each frame.

When forming the mel-spectrogram, each 0.975 second frame is further divided into frames of 25 milliseconds with an overlap of 15 milliseconds between adjacent frames, resulting in 96 frames. Hann window is applied to each frame followed by transformation to frequency domain using DFT of length 512 points. The DFT values are grouped using 64 mel-filters and the log of these results in a 96×64 mel-spectrogram. Illustration of the mel-spectrogram of the cough sounds from Fig. 1(a) is shown in Fig. 3(a).

The spectral characteristics of cough and non-cough sounds are learned from their mel-spectrogram representations using VGGish and YAMNet [9]. VGGish and YAMNet are pretrained CNNs for audio classification. These networks have been trained on Audio Set [13], a large YouTube audio dataset. VGGish is inspired by the popular VGG network for image classification [14] and YAMNet employs the MobileNets architecture [15].

The pretrained CNNs are combined with an RNN to learn the dependencies between a sequence of frames and predict if the recording is a cough recording or non-cough recording, that is, sequence-to-label classification. This work exploits BiLSTM, a type of RNN that uses two LSTMs: one to learn the sequences in the forward direction and another in the backward direction. This work uses two BiLSTM networks with 150 hidden units in each. The network weights are optimized using the adaptive moment estimation algorithm with an initial learn rate of 0.0003.

C. Baseline Method

Conventional classification methods are not suited to directly perform sequence-to-label classification. Therefore, for the baseline method, classifiers are first used to predict the probability of each 0.975 second frame to be cough or non-cough. The reference label for a frame is based on manual segmentation of the coughs. The probability of the recording being cough or non-cough is then determined as the *maximum* classification probability of the frames in the recording.

Mel-frequency cepstral coefficients (MFCCs) [12] are a widely used feature in audio classification. MFCCs are used as baseline features in this work. For computation of MFCCs, each 0.975 second frame is further divided into frames of 32 milliseconds with a 50% overlap. 13 MFCCs and their first and second derivatives are computed in each frame.

Binary classification is performed using the logistic regression (LR), random forest (RF), support vector machine (SVM), and RNN (BiLSTM) classifiers. While BiLSTM learns directly from the 39 MFCCs and their derivatives, for LR, RF, and SVM, the final feature vector for each 0.975 second frame is represented using the following 11 features: *mean*, *median*, *root mean square*, *maximum*, *minimum*, *1st and 3rd quartile*, *interquartile range*, *standard deviation*, *skewness*, and *kurtosis*. This results in a 429-dimensional MFCC feature set (including the first and second derivatives). These features are standardized using *z*-score and the discriminative features are identified using *t*-test (*p*-value of 0.05).

In addition, the pretrained network YAMNet has *Cough* as one of the prediction classes. We, therefore, use the pretrained YAMNet for classification of mel-spectrogram of frames and label a recording as cough if at least one frame is classified as cough. The pretrained VGGish could not be directly evaluated in this way as it has been trained for regression.

D. Evalution Metrics

The performance of the proposed and baseline methods is evaluated using sensitivity, specificity, accuracy, and area under the receiver operating characteristic (ROC) curve (AUC). Sensitivity and specificity are the proportion of cough and non-cough recordings that are correctly classified, respectively, and accuracy is the proportion of all recordings (cough and non-cough) that are correctly classified. The optimal threshold on the ROC curve is determined as the point on the curve that minimizes the distance to the point (0, 1).

III. Experimental Evaluation

A. Experimental Setup

The performance of the methods is evaluated in a 5-fold stratified cross-validation where 320 audio recordings (160 cough and 160 non-cough recordings) are used to train the classifier in each fold and 80 recordings (40 cough and 40 non-cough) are used to evaluate the performance of the classifier.

B. Classification Results

The cough vs. non-cough recording classification results using the baseline and proposed methods are given in Table II. An accuracy of 0.8700 (AUC=0.9422) is achieved using LR on the MFCC feature set. This drops to 0.8475 (AUC=0.9281) using SVM and improves to 0.9125 (AUC=0.9615) using BiLSTM. With an accuracy of 0.9350 (AUC=0.9816), the RF classifier outperforms LR, SVM, and BiLSTM on the MFCC feature set. This is a relative improvement of 2.47% to 10.32% in accuracy and 2.09% to 5.76% in AUC over these classifiers.

The classification accuracy further improves to 0.9550 (AUC=0.9871) using the pretrained YAMNet. This is a relative improvement of 2.14% in accuracy and 0.56% in AUC over the best results on the MFCC feature set, achieved by RF. Using the proposed CNN-RNN classification method, the accuracy improves to 0.9700 (AUC=0.9908) using YAMNet-BiLSTM and 0.9800 (AUC=0.9973) using VGGish-BiLSTM. As such, VGGish-BiLSTM achieves the highest classification accuracy of all the classification methods studied in this work. It produces a relative improvement of 4.81% in accuracy and 1.60% in AUC over the MFCC+RF classifier and a relative improvement of 2.62% in accuracy and 1.03% in AUC over the pretrained YAMNet. In Fig. 3(b), predictions of the VGGish network are investigated on mel-spectrogram of Fig. 3(a) using occlusion sensitivity [16], which suggests that it is focusing on the low frequency area in the cough frames.

While this work takes a frame-based approach for classification, the recordings are directly classified in [6]. We also experiment with this approach, computing the MFCC features across the entire recording and using LR, RF, SVM, and BiLSTM for classifying the recordings directly. This gives the highest accuracy of 0.8225 (AUC=0.9027), lower than all the frame-based methods considered in this work.

TABLE II. COUGH VS. NON-COUGH RECORDING CLASSIFICATION RESULTS USING THE BASELINE AND PROPOSED METHODS

Features/Input	Classifier	Sensitivity	Specificity	Accuracy	AUC
MFCC	LR	0.9000	0.8400	0.8700	0.9422
	RF	0.9350	0.9350	0.9350	0.9816
	SVM	0.8350	0.8600	0.8475	0.9281
	BiLSTM	0.9200	0.9050	0.9125	0.9615
Mel-Spectrogram	YAMNet (Pretrained)	0.9450	0.9650	0.9550	0.9871
	YAMNet-BiLSTM	**0.9800**	0.9600	0.9700	0.9908
	VGGish-BiLSTM	**0.9800**	**0.9800**	**0.9800**	**0.9973**

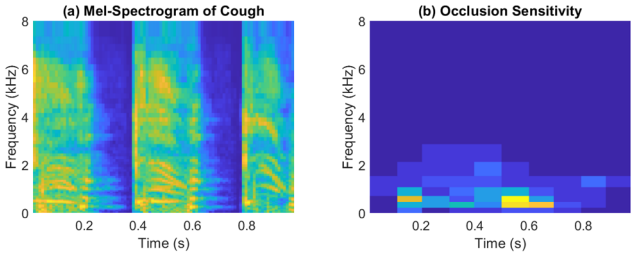

Fig. 3. (a) Mel-spectrogram of the cough sounds of Fig. 1(a) and (b) VGGish network visualizations using occlusion sensitivty.

Finally, the best performing classification model (VGGish-BiLSTM) is used to analyze the remaining 27,150 recordings in the COUGHVID dataset. These recordings have not been annotated; therefore, we only report the predictions (cough or non-cough) of the model which has shown strong classification performance on the annotated dataset of 400 recordings. The model predicts 8,871 (33%) recordings as non-cough recordings and 18,279 (67%) recordings as cough recordings.

IV. Discussion and Conclusion

A method for classification of cough vs. non-cough recordings is presented in this paper. The proposed method uses CNN to learn the spectral characteristics in the frames of audio recordings and RNN to learn the dependencies between the frame sequences. It achieves strong classification performance on a manually annotated dataset and outperforms various baseline methods, which include handcrafted features and conventional classifiers. In addition, the proposed method identifies a high proportion of unusable recordings in a crowdsourced dataset. While crowdsourced data can help in quickly collating large datasets for developing cough classification models, our findings highlight the need to exercise caution and rigorous data cleansing when using datasets sourced in this manner.

In a similar work [6], a cross-validation accuracy of 0.867 (AUC=0.964) is reported based on 68 handcrafted features, including MFCCs, and eXtreme Gradient Boosting classifier. However, there are important differences between our work and [6], such as a significantly smaller dataset (215 recordings) and the fact that they discarded cough recordings containing non-cough sounds. Some of the cough recordings in our work contain non-cough sounds as well, as depicted in Fig. 1(a), making it a more challenging task. In future, we plan to evaluate the proposed method on other crowdsourced datasets of cough recordings and also study cough sound quality.

References

[1] A. Murata, Y. Taniguchi, Y. Hashimoto, Y. Kaneko, Y. Takasaki, and S. Kudoh, "Discrimination of productive and non-productive cough by sound analysis," *Internal Medicine,* vol. 37, no. 9, pp. 732–735, 1998.

[2] 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.

[3] 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.

[4] C. Brown *et al.*, "Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 3474–3484.

[5] J. Laguarta, F. Hueto, and B. Subirana, "COVID-19 artificial intelligence diagnosis using only cough recordings," *IEEE Open Journal of Engineering in Medicine and Biology,* vol. 1, pp. 275–281, 2020.

[6] 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.

[7] H. Coppock, A. Gaskell, P. Tzirakis, A. Baird, L. Jones, and B. Schuller, "End-to-end convolutional neural network enables COVID-19 detection from breath and cough audio: a pilot study," *BMJ Innovations,* vol. 7, no. 2, pp. 356–362, 2021.

[8] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," in *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada, 2019, pp. 1–11.

[9] 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.

[10] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks,* vol. 18, no. 5, pp. 602–610, 2005.

[11] L. R. Rabiner and R. W. Schafer, *Theory and Applications of Digital Speech Processing*, First ed. New Jersey: Prentice Hall, 2011.

[12] 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.

[13] J. F. Gemmeke *et al.*, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, 2017, pp. 776–780.

[14] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," in *Proceedings of the 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, Kuala Lumpur, Malaysia, 2015, pp. 730–734.

[15] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *arXiv:1704.04861,* 2017.

[16] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Zurich, Switzerland, 2014, pp. 818–833.