Detecting Childhood Pneumonia Using Handcrafted and Deep Learning Cough Sound Features and Multilayer Perceptron*

Roneel V. Sharan[1], Kun Qian[2], and Yoshiharu Yamamoto[3]

***Abstract*— Pneumonia is one of the leading causes of morbidity and mortality in children. This is especially true in resource poor regions lacking diagnostic facilities, bringing about the need for rapid diagnostic tests for pneumonia. Cough is a common symptom of acute respiratory diseases, including pneumonia, and the sound of cough can be indicative of the pathological variations caused by respiratory infections. As such, in this paper we study objective cough sound evaluation for differentiating between pneumonia and other acute respiratory diseases. We use a dataset of 491 cough sounds from 173 children diagnosed either as having pneumonia or other acute respiratory diseases. We extract features which describe the temporal, spectral, and cepstral characteristics of the cough sound. These features are combined with feature embeddings from a pretrained deep learning network and used to train a multilayer perceptron for classification. The proposed method achieves a sensitivity and specificity of 84% and 73% respectively in differentiating between pneumonia and other acute respiratory diseases using cough sounds alone.**

I. INTRODUCTION

Pneumonia is the single largest infectious cause of death in children worldwide, accounting for 740,180 (14%) of all deaths in children under 5 years old in 2019 [1]. While pneumonia affects children all over the world, the vast majority of these deaths are in resource poor regions, such as southern Asia and sub-Saharan Africa. Indigenous children in developed countries are also disproportionately affected by pneumonia [2], [3].

The symptoms of pneumonia can include cough, breathing difficulty, fever, chest pain, amongst others. These symptoms can be used by the clinical algorithm developed by the World Health Organization to classify pneumonia in resource poor regions. However, other acute respiratory diseases can also cause similar symptoms, resulting in poor specificity of the algorithm and over prescription of antibiotics used for the treatment of pneumonia [4]. Chest radiography can be used for differential diagnosis [5] but it is not readily available in resource poor regions. This brings about the need for rapid diagnostic tests for pneumonia.

*Research supported in part by the Google Research Scholar Program, the Ministry of Science and Technology of the People's Republic of China with the STI2030-Major Projects 2021ZD0201900, the National Natural Science Foundation of China (No. 62272044), JSPS KAKENHI (No. 20H00569), the JST Mirai Program (No. 21473074), and the JST MOONSHOT Program (No. JPMJMS229B). *(Corresponding author: Roneel V. Sharan.)*

[1]Australian Institute of Health Innovation, Macquarie University, Sydney, NSW 2109, Australia (e-mail: roneel.sharan@mq.edu.au).

[2]School of Medical Technology, Beijing Institute of Technology, Beijing 10081, China (e-mail: qian@bit.edu.cn).

[3]Educational Physiology Laboratory, Graduate School of Education, The University of Tokyo, Japan (e-mail: yamamoto@p.u-tokyo.ac.jp).

Cough is a common symptom of acute respiratory diseases. Cough is comprised of three phases, inspiratory, compressive, and expiratory, and it is a vital defensive mechanism for lung health [6]. The sound of cough is associated with cough physiology. Different respiratory diseases can affect different part of the respiratory system. These pathological variations can be reflected in the sound of cough [7] and, therefore, be indicative of the respiratory disease [8], [9].

Despite its plausibility, the sound of cough has rarely been studied in differentiating pneumonia from other acute respiratory diseases. Earlier works [4], [10] in detecting childhood pneumonia using cough sound analysis use conventional feature engineering and machine learning techniques, such as the use of handcrafted features and logistic regression classification. In addition, they employ a small number of subjects and the cough classification model of [11], [12] is reliant on clinical symptoms.

In this work, we propose a method to detect childhood pneumonia using only cough sounds. Similar to [4], [10], [11], [12], our work makes use of various handcrafted features, capturing different characteristics of the cough sound. In addition, our work makes the following contributions when compared to these earlier works. Firstly, we use a set of deep learning features extracted from a pretrained audio classification network. Secondly, we train a multilayer perceptron on the combined handcrafted and deep learning feature set to differentiate between pneumonia and non-pneumonia cough sounds. Multilayer perceptron is a fully connected class of feedforward artificial neural network that has the ability to learn complex relationships between the input features so that they can be combined into higher-order representations. The proposed method is evaluated on a clinically verified dataset of cough sounds from children diagnosed as having pneumonia or other acute respiratory diseases. The dataset has almost twice as many subjects compared to [4], [10].

II. MATERIALS AND METHODS

A. Dataset

In this work, we use a dataset of cough sound recordings collected at West China Second University Hospital of Sichuan University [13]. An overview of the dataset is provided in Table I. The dataset has audio recordings of cough sounds from 173 children with acute respiratory diseases which can be grouped into two classes: *pneumonia* and *non-pneumonia*. The pneumonia class has 82 subjects (43 male and 39 female) of which 55 subjects are diagnosed as having pneumonia, 23 subjects as bronchopneumonia, and

TABLE I
OVERVIEW OF THE DATASET USED IN THIS WORK

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

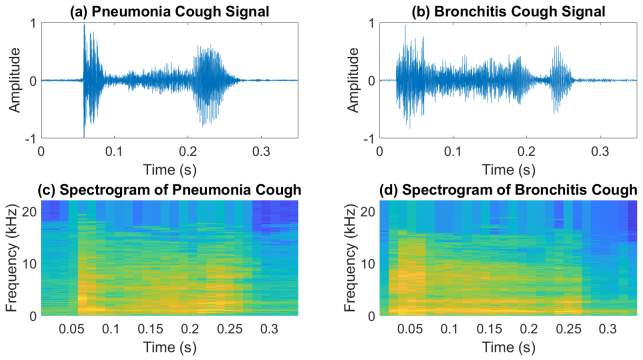

Fig. 1. Waveform of cough sounds for (a) pneumonia and (b) bronchitis, and their spectrogram representation showing the frequency characteristics in (c) and (d) respectively.

4 subjects as lobar pneumonia. The non-pneumonia class has 91 subjects (51 male and 40 female) of which 80 subjects have acute bronchitis, 6 subjects have acute bronchiolitis, and 5 subjects have acute asthmatic bronchitis. The disease diagnosis was done according to [14]. The children are aged 0 to 11 years with majority aged one year or less.

The audio recordings are available in the MP3 file format at a sampling frequency of 44.1 kHz. The recordings are made in a hospital environment and contain background noises, such as speech and sounds from medical devices. The total duration of pneumonia recordings is 372.10 seconds and 320.61 seconds for non-pneumonia recordings. All the recordings were manually screened for use in this work. Two pneumonia recordings were excluded as they were determined not to contain any cough sound. Another pneumonia recording was excluded because it could not be determined if the respiratory sounds from the infant were cough or non-cough. Each of the remaining recordings have one or more cough sounds which are manually segmented for analysis in this work. This gives us a total of 268 cough sounds in the pneumonia class and 223 cough sounds in the non-pneumonia class. Illustration of pneumonia and non-pneumonia (bronchitis) cough waveforms along with their spectrogram representations are given in Fig. 1.

B. Proposed Method

An overview of the proposed method in classifying pneumonia and non-pneumonia cough sounds is illustrated in Fig. 2. All recordings are converted to the WAV file format and the cough signals in each recording are manually segmented to determine the start and end point of each cough. The manually segmented cough signals are processed to extract two sets of features [15]: a set of *handcrafted features* and a set of transfer learning-driven *deep learning features*.

Handcrafted Features: This work utilizes two subsets of handcrafted features: *cepstral* and *temporal and spectral*. In computing the handcrafted features, each cough signal is divided into frames of 25 milliseconds with an overlap of 15 milliseconds between adjacent frames. The cepstral features are *mel-frequency cepstral coefficients (MFCCs)* [16], a widely used feature in audio classification tasks that utilizes frequency scales based on the auditory perception. We compute 13 MFCCs and the first and second derivatives of these coefficients [17] in each frame, resulting in a matrix of size $39 \times n_c$ for each cough, where n_c is the number of frames in the c^{th} cough. These raw features are represented using the *mean* and *standard deviation* statistical measures. If the recording has only one cough, these statistics are computed across all the frames in the cough. If the recording has multiple coughs, these statistics are computed across all the frames from all the coughs. These result in a 78-dimensional MFCC feature subset for each recording.

The second handcrafted feature subset has 15 features which capture the temporal and spectral characteristics of the cough signal, computed in each frame similar to MFCCs. These are the *zero-crossing rate*, *short-time energy*, *spectral centroid*, *spectral crest*, *spectral decrease*, *spectral entropy*, *spectral flatness*, *spectral flux*, *spectral kurtosis*, *spectral roll-off point*, *spectral skewness*, *spectral slope*, *spectral spread*, *pitch*, and *harmonic ratio* [18], [19]. These are once again represented using the mean and standard deviation statistical measures, resulting in a 30-dimensional temporal and spectral feature subset for each recording.

Deep Learning Features: The deep learning feature set has 128 *VGGish* feature embeddings from each cough signal. These are extracted using a pretrained convolutional neural network for audio classification [20]. The VGGish is inspired by the popular VGG networks in image classification. The VGGish has been trained on a large YouTube audio dataset of 128-dimensional embeddings. In computing the VGGish features, each cough signal is zero-padded or cropped to 0.975 seconds and transformed into a 94×64 log mel-spectrogram. The mel-spectrogram time-frequency representation forms input to the VGGish network for extracting the feature embeddings. In the event a recording contains multiple coughs, the feature embeddings are averaged across the coughs.

The combined feature vector is, therefore, 236-dimensional (78 MFCC features, 30 temporal and spectral features, and 128 VGGish features). These cough features are used for binary classification (pneumonia vs non-pneumonia) using the following classifiers: *random forest (RF)*, *support vector machine (SVM)*, and *multilayer perceptron (MLP)*. For the RF and SVM classifiers, the

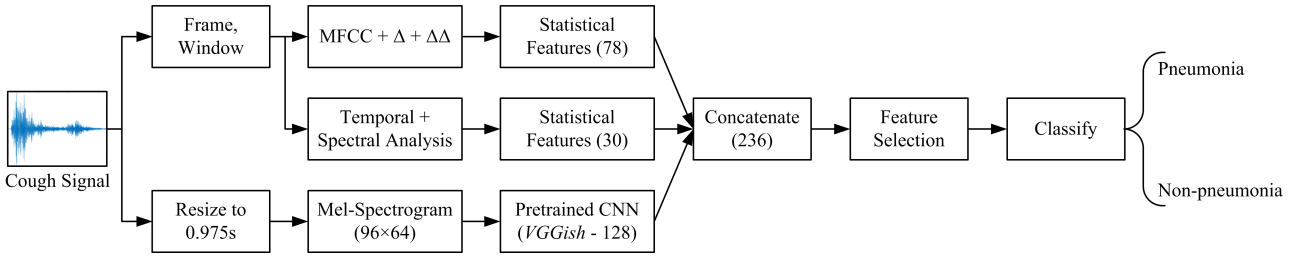

Fig. 2. Overview of the proposed method in pneumonia vs non-pneumonia cough sound classification.

discriminative features are identified using *t*-test and elastic net [21], while the full feature vector is used as input to the MLP classifier. The MLP has two hidden layers. Each hidden layer has 256 neurons and the rectified linear unit activation function. The network is trained using adaptive moment estimation.

C. Evaluation Metrics

The classification performance is measured using sensitivity, specificity, accuracy, and F_1 score computed as

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

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

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

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

where *TP* is the number of true positives, *TN* is the number of true negatives, *FP* is the number of false positives, and *FN* is the number of false negatives. The positive and negative classes are pneumonia and non-pneumonia, respectively. The optimal threshold on the ROC curve is selected such that sensitivity is greater than specificity by at least 10 percentage points, that is, prioritizing detection of pneumonia over non-pneumonia. The area under the curve (AUC) of the receiver operating characteristic (ROC) curve is also used, as a single measure of classification performance.

III. EXPERIMENTAL EVALUATION

A. Experimental Setup

The performance of the proposed pneumonia vs non-pneumonia cough sound classification method is evaluated in leave-one-out cross-validation whereby features from one subject are used for testing and the features from the remaining subjects are used for training. This procedure is repeated such that each subject is used to test the classification model once. In each fold, the features are normalized using *z*-score normalization. For the RF and SVM classifiers, we present results using the feature selection method that produced the highest overall results. With the *t*-test, the discriminative features are selected using a *p*-value threshold of 0.05. With the elastic net, the discriminative features are selected using the minimum cross-validated mean square error. For both feature selection methods, the discriminative features in each fold are identified on the training data. The results for all the classifiers are presented using the handcrafted feature set, the deep learning feature set, and the combined feature set.

B. Cough Classification Results

The results for pneumonia vs non-pneumonia cough sound classification are presented in Table II. The RF classifier achieves best results on the handcrafted and deep learning feature sets when the features are selected using *t*-test while the elastic net method of feature selection yields the best results on the combined feature set. An accuracy of 0.6647 and F_1 of 0.6705 is achieved with the handcrafted features. With an accuracy of 0.6529 and F_1 of 0.6550, there is a slight drop in the classification performance with the DL features. However, with an accuracy of 0.6882 and F_1 of 0.6901, the best classification results are achieved using the combined feature set.

With SVM classification, the *t*-test method of feature selection produces the best results on all three feature sets. Similar to the trend with the RF classifier, the results using the deep learning features are slightly lower than using handcrafted features. However, with an accuracy and F_1 of 0.7059, the best results using SVM are with the combined feature set, same as what is observed using the RF classifier.

All performance metrics are seen to improve when using the MLP classifier. On the handcrafted feature set, the MLP classifier achieves an accuracy of 0.7412, a relative improvement of 11.51% over RF and 7.70% over SVM, and F_1 of 0.7412, a relative improvement of 10.54% over RF and 7.40% over SVM. On the deep learning feature set, the MLP classifier achieves an accuracy of 0.6882, a relative improvement of 5.41% over RF and 7.33% over SVM, and F_1 of 0.6901, a relative improvement of 5.36% over RF and 7.27% over SVM. On the combined feature set, the accuracy using MLP is 0.7765, a relative improvement of 12.83% over RF and 10.00% over SVM, and F_1 is 0.7765, a relative improvement of 12.52% over RF and 10.00% over SVM. As such, the MLP classifier outperforms the RF and SVM classifiers and, once again, the best classification results are achieved on the combined feature set.

Box plot of the most significant feature (lowest *p*-value using *t*-test) from the handcrafted feature set and deep learning feature set are shown in Fig. 3. VGGish feature embedding 94 is determined to be the most significant feature

TABLE II
PNEUMONIA VS NON-PNEUMONIA COUGH SOUND CLASSIFICATION RESULTS

Feature Set	Feature Selection Method	Classifier	Classification Results				
			Sensitivity	Specificity	Accuracy	AUC	F_1
Handcrafted features	T-Test	RF	0.7342	0.6044	0.6647	0.6769	0.6705
DL features	T-Test		0.7089	0.6044	0.6529	0.7251	0.6550
Handcrafted + DL features	Elastic Net		0.7468	0.6374	0.6882	0.7426	0.6901
Handcrafted features	T-Test	SVM	0.7468	0.6374	0.6882	0.7325	0.6901
DL features	T-Test		0.6962	0.5934	0.6412	0.7068	0.6433
Handcrafted + DL features	T-Test		0.7595	0.6593	0.7059	0.7749	0.7059
Handcrafted features	–	MLP	0.7975	0.6923	0.7412	0.8064	0.7412
DL features	–		0.7468	0.6374	0.6882	0.7311	0.6901
Handcrafted + DL features	–		**0.8354**	**0.7253**	**0.7765**	**0.8208**	**0.7765**

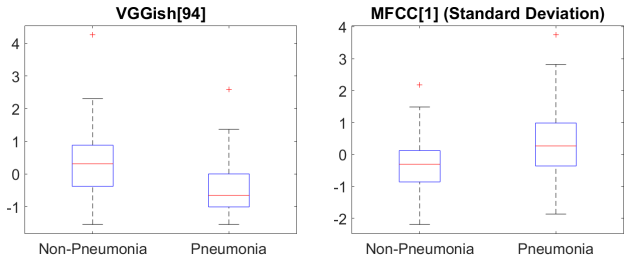

Fig. 3. Box plot of the most significant feature (lowest p-value) from each feature set.

followed by the standard deviation of the 1st mel-frequency cepstral coefficient.

IV. CONCLUSION

This work proposes a method for classifying pneumonia and non-pneumonia cough sounds using handcrafted and deep learning features, and MLP. With a sensitivity of 0.8354, specificity of 0.7253, accuracy of 0.7765, AUC of 0.8208, and F_1 of 0.7765, these are the best results of all the feature sets and classifiers considered in this work. Our work, however, has some limitations. While our dataset has a greater number of subjects than other similar works, such as [4], [10], the non-pneumonia group is primarily comprised of bronchitis subjects. In the future, we plan to evaluate our method with more cough recordings from other pediatric acute respiratory diseases. In addition, the cough sounds in this work are manually segmented. In the future, we aim to evaluate our algorithms with automatically segmented cough sounds, such as using neural networks [8], [15]. Automatic cough segmentation is an important step in achieving a fully automated cough sound-based pneumonia detection system.

REFERENCES

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

[2] T. Kovesi, "Respiratory disease in Canadian First Nations and Inuit children," *Paediatr. Child Health*, vol. 17, no. 7, pp. 376–380, 2012.

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

[4] U. R. Abeyratne, V. Swarnkar, A. Setyati, and R. Triasih, "Cough sound analysis can rapidly diagnose childhood pneumonia," *Ann. Biomed. Eng.*, vol. 41, no. 11, pp. 2448–2462, 2013.

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

[6] A. B. Chang, "The physiology of cough," *Paediatr. Respir. Rev.*, vol. 7, no. 1, pp. 2–8, 2006.

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

[8] R. V. Sharan, U. R. Abeyratne, V. R. Swarnkar, and P. Porter, "Automatic croup diagnosis using cough sound recognition," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 2, pp. 485–495, 2019.

[9] 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," *Biomed. Signal Process. Control*, vol. 68, Art. no. 102722, 2021.

[10] K. Kosasih, U. R. Abeyratne, V. Swarnkar, and R. Triasih, "Wavelet augmented cough analysis for rapid childhood pneumonia diagnosis," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 4, pp. 1185–1194, 2015.

[11] P. Porter *et al.*, "A prospective multicentre study testing the diagnostic accuracy of an automated cough sound centred analytic system for the identification of common respiratory disorders in children," *Respir. Res.*, vol. 20, no. 1, Art. no. 81, 2019.

[12] P. P. Moschovis *et al.*, "A cough analysis smartphone application for diagnosis of acute respiratory illnesses in children," *Am. J. Respir. Crit. Care Med.*, vol. 199, p. A1181, 2019.

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

[14] Y. Hu and Z. F. Jiang, *Zhu Fu Tang Practical Pediatrics*, 8th ed. Beijing: People's Health Publishing House, 2015.

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

[16] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Audio Speech Lang. Process.*, vol. 28, no. 4, pp. 357–366, 1980.

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

[18] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the CUIDADO project," *IRCAM*, Paris, France, Technical Report 2004.

[19] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1997, pp. 1331–1334.

[20] S. Hershey *et al.*, "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2017, pp. 131–135.

[21] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Stat. Soc. B Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, 2005.