INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

DeepLung: Smartphone Convolutional Neural Network-based Inference of Lung Anomalies for Pulmonary Patients

Mohsin Y Ahmed[1]*, *Md Mahbubur Rahman*[2], *Jilong Kuang*[2]

[1]University of Virginia
[2]Samsung Research America

mohsin.ahmed@cs.virginia.edu, m.rahman2@samsung.com, jilong.kuang@samsung.com

Abstract

DeepLung is an end-to-end deep learning based audio sensing and classification framework for lung anomaly (e.g. cough, wheeze) detection for pulmonary patients from streaming audio and inertial sensor data from a chest-held smartphone. We design and develop 1-D and 2-D convolutional neural networks for DeepLung, and train them using the Interspeech 2010 Paralinguistic Challenge features. Two different audio windowing schemes: i) real-time respiration cycle based natural windowing, and ii) static length windowing are compared and experimented with. Classifiers are developed considering 2 different system architectures: i) mobile-cloud hybrid architecture, and ii) mobile in-situ architecture. Patient privacy is preserved in the phone by filtering speech with a shallow classifier. To evaluate DeepLung, a novel and rigorous lung activity dataset is made by collecting audio and inertial sensor data from more than 131 real pulmonary patients and healthy subjects and annotated accurately by professional crowdsourcing. Experimental results show that the best combination of DeepLung convolutional neural network is 15-27% more accurate when compared to a state-of-the-art smartphone based body sound detection system, with a best F1 score of 98%.

Index Terms: pulmonary diseases, smartphone, convolutional neural network

1. Introduction

Chronic pulmonary diseases like asthma and chronic obstructive pulmonary disease (COPD) are one of the leading cause of death worldwide, causing nearly 7% of all deaths in the United States in recent years [1]. An estimated 40 million people in USA are suffering from these diseases [2, 3], causing more than $130 billion in health-care costs [4].

A common syndrome of chronic pulmonary diseases is the obstruction of the respiratory airway. Therefore COPD exacerbation or asthma attacks are often accompanied by lung originated anomalies like wheezing, rhonchus, crackles and coughing. With the increasing usage of smartphones among the patient population and wide availability of smartphones equipped with hardware capable of running advanced real-time signal processing and shallow/deep learning algorithms, it is possible to use them as a home health-care disease monitoring tool for pulmonary patients. However, such pulmonary anomalous audio sensing by smartphone is concerning due to possibility of infringement of patient privacy. Also, well annotated high quality pulmonary activity data collected by smartphone is a scarce entity to date to make usable machine learning models.

In this paper, we present DeepLung which is an end-to-end deep learning audio sensing and classification framework for lung anomaly (e.g. cough, wheeze) detection for pulmonary patients. During usage, a pulmonary patient holds the phone, with DeepLung installed, by his chest for several minutes. DeepLung listens to the internal body sounds using the built-in phone microphone and fine trained convolutional neural network (CNN) either in the phone or in the cloud detects these unusual lung sounds. DeepLung can adopt either a dynamic respiration cycle based physiological and natural windowing of audio, or a static length windowing. Respiration cycle is detected from the expansion and contraction of the patient's chest from the phone's inertial sensors' readings. DeepLung uses the Interspeech 2010 Paralinguistic Challenge features [5] as input to it's 1-D and 2-D CNNs for classifying lung and other confounding body sounds. DeepLung classifiers have been designed considering 2 possible system architectures: i) a mobile-cloud hybrid architecture where the phone captures audio, does windowing, preprocessing and speech filtering, and then the lung sounds are classified in the cloud, and ii) a mobile in-situ architecture where all processing is done in the phone. DeepLung ensures patient privacy in the mobile-cloud architecture by filtering out speech frames in the local phone with high accuracy using a shallow classifier.

To tackle the challenge of limited available pulmonary activity data to make appropriate deep learning models, we collected audio and inertial sensor data using a smartphone from 131 pulmonary patients and healthy subjects over controlled and well designed pulmonary activities. We collaborated with a crowdsourcing company to ensure high quality annotation of different pulmonary anomalous events in this novel dataset. A series of experiments were performed using this novel dataset and performance was compared with the BodyBeat [6] system.

Several other shallow and deep learning based pulmonary activity detection works like wheeze detection [7–10, 10–15] and cough detection [16–18] exist in literature. However, they often use limited training data which is not collected with a commodity smartphone. Our CNN models are built using the first ever pulmonary activity data collected from 131 real patients and healthy subjects using smartphones, and trained using the Interspeech 2010 Paralinguistic Challenge features which are tailored to detect non speech human sounds.

The contributions of this paper are: 1) designing and comparing 1-D and 2-D CNNs for pulmonary anomaly detection using Interspeech 2010 Paralinguistic Challenge acoustic features, 2) experimental results verified by the first, extensive pulmonary sound dataset collected with commodity smartphones from 131 real patients and healthy subjects, and 3) comparison of lung sound detection performance between respiration cycle based windowing and static sized windowing on this novel dataset.

[*] The subject matter of this paper was developed while the author was doing internship at Samsung Research America.

http://dx.doi.org/10.21437/Interspeech.2019-2953

2. Data preparation

2.1. Pulmonary data collection

Real-world pulmonary anomalous activity audio data was collected from 131 patients and healthy subjects using a Samsung Galaxy Note 8 smartphone. The subjects were asked to hold the phone by their chest and perform 7 different pulmonary activities during which audio and inertial sensor data were recorded by the phone. There were total 91 chronic pulmonary patients and 40 healthy people among the subjects. 69 of them were asthma patients, 9 were COPD patients, and 13 exhibited comorbidity of both asthma and COPD. A Zephyr bio-harness was worn by the subjects during data collection for respiration ground truth. On average, there was around 40 minutes of continuous audio and inertial sensor data collected per subject.

The data collection protocol consisted of each subject performing the following tasks while holding the phone by their chest: 1) pulmonary function test (3 efforts), 2) cough naturally/voluntarily for 2 minutes, 3) making A-vowel ('Aaaa...') sound for as long as possible, 4) speaking freely on any topic for 3-5 minutes, 5) reading a neutral paragraph for 3-5 minutes, 6) sit silently for 1 minute and count breaths, 7) lying down silently for 1 minute and count breaths. These 7 tasks were chosen carefully by expert researchers to make the dataset with high quality and enough variability in terms of disease specific pulmonary anomalies.

2.2. Data annotation

Due to the possible challenge and errors in annotating such a large pulmonary activity dataset, we collaborated with a crowdsourcing company to perform non-expert annotation of the dataset. The annotators used an interface to mark the start and end of each cough, wheeze, speech, throat clearing, abdominal sounds, and other confounding sounds. A section in the interface provided training to the annotators with examples of each pulmonary event, and how to label them in the audio. A second set of annotators reviewed and corrected possible incorrect annotations during the first phase. Some part of these annotations were also reviewed and verified by expert researchers in this domain to ensure a rich and high quality pulmonary activity dataset. In total, the dataset contained more than 3,500 instances of cough, wheeze, throat-clearing, abdominal sounds, and other confounding noises, and more than 42,000 instances of speech. To the best of our knowledge, this is the first ever pulmonary disease specific anomalous activity dataset collected solely with commodity smartphone sensors without using any additional customized microphone from such a large number of real pulmonary patients.

2.3. Data normalization

The volume of each raw audio frame in the dataset was normalized to 1 by dividing the whole frame by the frame peak value. This was done to eliminate any bias due to frame volume for the classifiers.

3. Methods

DeepLung, when held by the chest by a patient as in Figure 1, listens to the lung and confounding body sounds and detects pulmonary anomalies like cough and wheeze, and other confounding body sounds (like throat clearing and abdominal sounds).

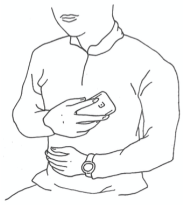

Figure 1: *Example usage of DeepLung by a pulmonary patient.*

3.1. Architecture

We design and develop DeepLung algorithms considering 2 different implementation architectures as follows:

3.1.1. Mobile-cloud hybrid processing

In this architecture, the phone normalizes and windows the audio stream coming from the patient's body, and filters out speech frames using a shallow in-phone classifier for patient privacy. Then it uploads the body sound frames to the cloud where a CNN does lung sound (cough and wheeze) classification. For handling confounding body sounds (throat clearing, abdominal sounds) there is an additional class, hence the CNN in the cloud does a 3-class classification: cough, wheeze, and other.

3.1.2. Mobile in-situ processing

In this architecture, the mobile does in-situ end-to-end audio processing, and the sound frames never leave the device (as opposed to uploading to the cloud). An offline trained CNN in the phone does a 4-class classification: speech, cough, wheeze, and other.

3.2. Audio windowing

We experimented with 2 different audio windowing schemes for DeepLung as following:

3.2.1. Respiration cycle based natural windowing

Since DeepLung operates when a patient holds the phone by the chest, the respiratory cycle can be detected in real-time from the expansion and contraction of the chest using the phone's inertial sensors [19], and the audio stream can be windowed by the respiratory cycle for classification. Such respiratory cycle based natural windowing ensures that all the distinct phases of a cough and a wheeze resides in a single window.

3.2.2. Static windowing

We also experimented with static fixed size windows of 60 ms length and 10 ms shift to extract low level acoustic features to train the CNN. Such static windowing results in a much larger amount of training instances compared to respiration based natural windowing.

3.3. In-phone speech filtering

In the mobile-cloud hybrid processing architecture, DeepLung uses a shallow support vector machine (SVM) classifier to separate and discard speech frames from being uploaded to the cloud for patient privacy. Because of the complex physiological generation process, lung and other in-body sounds are situated at a much lower region (20-300 Hz) of frequency spectrum than speech (300-3500 Hz) [6], and hence have high separability. A Similar approach has been shown to separate speech from internal body sounds with high accuracy using limited number of acoustic features [20].

3.4. Feature extraction

We used the Interspeech 2010 Paralinguistic Challenge feature set [5] to extract a total of 1582 acoustic features from the normalized audio frames using the OpenSmile [21] tool. The tool first extracts 38 acoustic low-level descriptors (LLD) related to energy, pitch, spectral, cepstral, mel-frequency, voicing probability, shimmer and jitter, and their first order delta regression coefficients. Then 21 different statistical functionals are applied to the LLDs to map a audio time series signal of variable length (based on respiration window duration) into a static sized (1582) feature vector. For respiration cycle based windowing of audio, we used all the 1582 features, while for static sized 60 ms window, we used only the 38 LLDs as features. Table 1 summarizes the features.

3.5. Deep neural networks structure

For our analysis to classify the lung sounds, we used both 1-dimensional and 2-dimensional CNN models.

3.5.1. 1-D CNN models

For respiration cycle windowed audio represented as 1582 features, we used a CNN with 1 convolutional layer and 1 max pooling layer. Instead of adopting a fully-connected layer for classification, the classifier outputs the category confidence via the max pooling layer, and then the resulting vector is fed into the output (softmax) layer [22]. Since, respiration cycle based samples are comparatively fewer in number, we did not use a fully connected layer to reduce the number of parameters and reduce chances of overfitting [23]. For static windowed audio represented as 38 features, we used a CNN with 2 convolutional layers, 1 max pooling layer, and a fully connected layer, as samples are much larger in quantity (more than 800,000).

3.5.2. 2-D CNN models

We transformed the 1-D feature vectors into 2-D by transposing it and taking a dot product with the original feature vector. Since the feature space for respiration cycle windowed audio is high (1582), we reduced it's dimension by choosing the best 50 features using mutual information before transforming it to a 50x50 dimension 2-D feature vector to reduce memory consumption. The static windowed 1-D features were transformed to 38x38 size 2-D vectors similarly. Then we use them to train 2-D CNNs similar in structure to the 1-D CNNs. Table 2 describes the structures of the different CNNs we used to make DeepLung models.

4. Experimental results

4.1. Training amount

We evaluate the F1 scores of our different CNNs for different amounts of training. For the dataset windowed by respiration cycles, we vary the training amount between 50-1000, and for the static windowed dataset, we vary the training amount between 10,000-200,000. We use 80% of the dataset for training, and remaining 20% for validation. From the training set, we randomly select the target amount of samples to make the classification model, and test on the validation set. Each experiment is repeated 5-10 times.

Table 1: *Low level descriptors and high level functionals used for CNN training*

Low Level Descriptors	Functionals
PCM loudness, MFCC (0-14), Log Mel Frequency Band [0-7], Line Spectral Pair Frequency [0-7], Fundamental Frequency, Fundamental Frequency Envelope, Voicing Probability, Jitter local, Jitter consecutive frame pairs, Shimmer local	position maximum/minimum, arithmetic mean, standard deviation, skewness, kurtosis, linear regression coefficient (1st, 2nd), linear regression error (quadratic, absolute), quartile (1st, 2nd, 3rd), quartile range 2-1/3-2/3-1, percentile (1st, 99th), percentile range 99-1, up-level time 75/90

The results are presented in Table 3. We discuss our observations in the following sections:

4.1.1. 1-D CNN vs. 2-D CNN

It is noted from Table 3 that 2-D CNN performs really poorly compared to 1-D CNN for respiration cycle windowed audio. The reason is that, we reduced the dimensionality of the feature vector from 1582 to 50 by mutual information based feature selection for memory efficiency before transforming it to 2-D, which cost loss of information leading to poor performance. For static windowed audio, the 1-D and 2-D CNN performances are close as no dimension reduction was done on the original feature space. For both cases, f-1 scores increase with increase in training samples.

4.1.2. Natural windowing vs. static windowing

Respiration cycle based natural windowing of lung sounds performs between 4-20% better than static windowed lung sounds, as seen from Table 3. Respiration based physiological and natural windowing ensures that the distinct phases of lung anomalous sounds resides in a single frame, and hence have better separability. However, we hypothesize that with a deeper and better neural network design, it is possible to further improve the performance of static windowed classifiers.

4.1.3. 3-class vs. 4-class classification

Performances of both 3-class and 4-class CNNs are similar when a larger amount of training is available. 4-class models have an additional class of speech, which is easily separable from lung and other body sounds due to its position in a higher frequency spectrum. 4-class CNN performance is a little poorer than 3-class for respiration windowed audio when limited training is provided, but it becomes close to the 3-class models as additional training samples are provided.

4.2. Leave-1-person-out test

This experiment was conducted considering the realistic scenario when a patient tries to use a DeepLung system which has been trained by speech and lung sound samples from other people. This experiment was done on the 3-class data using 1-D CNN. For this experiment, we randomly picked 80% sam-

Table 2: *Structure of different CNN models*

Network	Detail
1-D CNN for respiration windowed audio	conv1: 8 filters, kernel size 3, 1 stride pooling: pool_size 2, 1 stride classification layer: softmax
1-D CNN for static windowed audio	conv1: 8 filters, kernel size 3, 1 stride conv2: 8 filters, kernel size 3, 1 stride pooling: pool_size 2, 1 stride fully connected layer: 100 neurons classification layer: softmax
2-D CNN for respiration windowed audio	conv1: 8 filters, kernel size 3x3, 1x1 stride pooling: pool_size 2x2, 1x1 stride classification layer: softmax
2-D CNN for static windowed audio	conv1: 8 filters, kernel size 3x3, 1x1 stride conv2: 8 filters, kernel size 3x3, 1x1 stride pooling: pool_size 2x2, 1x1 stride fully connected layer: 100 neurons classification layer: softmax

Table 3: *F1 scores of different CNN models of DeepLung*

Respiration windowed					Static windowed				
	3-class classification (in cloud)		4-class classification (in phone)			3-class classification (in cloud)		4-class classification (in phone)	
# of training samples	**1-D CNN**	**2-D CNN**	**1-D CNN**	**2-D CNN**	**# of training samples**	**1-D CNN**	**2-D CNN**	**1-D CNN**	**2-D CNN**
50	.80±.08	.63±.251	.72±.114	.52±.267	**10,000**	.65±.034	.71±.038	.68±.037	.69±.03
100	.92±.035	.70±.284	.84±.06	.48±.3	**50,000**	.72±.028	.76±.02	.77±.017	.78±.022
500	.96±.012	.34±.137	.96±.009	.68±.384	**100,000**	.79±.015	.80±.009	.81±.009	.80±.01
1000	.98±.002	.60±.296	.98±.004	.97±.011	**200,000**	.82±.012	.82±.01	.84±.005	.82±.01

Table 4: *DeepLung 1-D CNN F1 score comparison with BodyBeat.*

	# of training samples		
	100	**500**	**1000**
DeepLung	.924±.035	.967±.012	.983±.002
BodyBeat	.65±.097	.80±.024	.83±.013

ples from each of the 2 datasets (respiration and static windowed), divided these samples to their corresponding 131 subjects, trained the CNN on all-but-1 subject, and tested on the samples from the remaining subject. This was done for each of the 131 subjects, with the experiment being repeated 5-10 times. The average F1 score (.995±.025) for respiration windowed audio was better than static windowed audio (.796±.15) for this test.

4.3. Comparison with BodyBeat

We compare DeepLung 1-D respiration windowed CNN performance with the BodyBeat [6] classifier, which is a mobile based body sound detection system using specialized microphone hardware. We implemented BodyBeat's algorithm with a linear discriminant classifier with 30 best features chosen using mutual information based feature selection from the 1582 original features. An 80-20% random split was done on the respiration windowed dataset to make the training and validation sets. Both classifiers were trained using same amount of training from the training set and tested on the validation set. Table 4 shows that DeepLung is 15-27% better than BodyBeat for various amount of training, although using only commodity smartphone hardware (as opposed to external microphone).

5. Discussion and conclusion

We primarily focused on detecting pulmonary anomalies like coughing and wheezing. Aside from these, there are other medically defined pulmonary disease symptoms like rhonchus and crackles. As more of these sounds become available, models for detecting these can be developed with our existing CNNs. There is scope to further engineer our 2-D CNNs for further improvement of performance. As deep neural networks have been demonstrated to be more noise resilient and location independent [24] compared to shallow learners, DeepLung is more realistic to use in a real world environment.

To summarize, we present DeepLung which is an end-to-end deep learning based audio sensing and classification framework for lung anomaly (like cough, wheeze) detection for pulmonary patients from streaming audio and inertial sensor data from a chest-held smartphone. As more and more pulmonary activity data becomes available, DeepLung will be an emerging technology enabling caregivers to do longitudinal daily remote monitoring of their patients, and analyze the effect of a particular prescribed medication over time. Finally, DeepLung is cost effective and user friendly as no external microphone is needed.

6. References

[1] "Webmd," https://www.webmd.com/lung/copd/news/20170929/respiratory-disease-death-rates-have-soared.

[2] "Cdc asthma," https://www.cdc.gov/nchs/fastats/asthma.htm.

[3] "Cdc copd," https://www.cdc.gov/nchs/fastats/copd.htm.

[4] A. J. Guarascio, S. M. Ray, C. K. Finch, and T. H. Self, "The clinical and economic burden of chronic obstructive pulmonary disease in the usa," *ClinicoEconomics and outcomes research: CEOR*, vol. 5, p. 235, 2013.

[5] B. Schuller et al, "The interspeech 2010 paralinguistic challenge," in *INTERSPEECH 2010*, 2010, pp. 2794–2797.

[6] T. Rahman et al, "Bodybeat: a mobile system for sensing non-speech body sounds." in *MobiSys*, vol. 14, 2014, pp. 2–13.

[7] R. Riella, P. Nohama, and J. Maia, "Method for automatic detection of wheezing in lung sounds," *Brazilian Journal of Medical and Biological Research*, vol. 42, no. 7, pp. 674–684, 2009.

[8] N. Sengupta, M. Sahidullah, and G. Saha, "Lung sound classification using cepstral-based statistical features," *Computers in biology and medicine*, vol. 75, pp. 118–129, 2016.

[9] G. D. Sosa, A. Cruz-Roa, and F. A. González, "Automatic detection of wheezes by evaluation of multiple acoustic feature extraction methods and c-weighted svm," in *10th International Symposium on Medical Information Processing and Analysis*, vol. 9287. International Society for Optics and Photonics, 2015, p. 928709.

[10] H.-K. Ra, A. Salekin, H. J. Yoon, J. Kim, S. Nirjon, D. J. Stone, S. Kim, J.-M. Lee, S. H. Son, and J. A. Stankovic, "Asthmaguide: an asthma monitoring and advice ecosystem," in *2016 IEEE Wireless Health (WH)*. IEEE, 2016, pp. 1–8.

[11] P. Bokov, B. Mahut, P. Flaud, and C. Delclaux, "Wheezing recognition algorithm using recordings of respiratory sounds at the mouth in a pediatric population," *Computers in biology and medicine*, vol. 70, pp. 40–50, 2016.

[12] L. Mendes, I. Vogiatzis, E. Perantoni, E. Kaimakamis, I. Chouvarda, N. Maglaveras, V. Tsara, C. Teixeira, P. Carvalho, J. Henriques *et al.*, "Detection of wheezes using their signature in the spectrogram space and musical features," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE, 2015, pp. 5581–5584.

[13] N. Nakamura, M. Yamashita, and S. Matsunaga, "Detection of patients considering observation frequency of continuous and discontinuous adventitious sounds in lung sounds," in *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the*. IEEE, 2016, pp. 3457–3460.

[14] M. Himeshima, M. Yamashita, S. Matsunaga, and S. Miyahara, "Detection of abnormal lung sounds taking into account duration distribution for adventitious sounds," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*. IEEE, 2012, pp. 1821–1825.

[15] S. Matsunaga, K. Yamauchi, M. Yamashita, and S. Miyahara, "Classification between normal and abnormal respiratory sounds based on maximum likelihood approach," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 517–520.

[16] S. Matos, S. S. Birring, I. D. Pavord, and D. H. Evans, "An automated system for 24-h monitoring of cough frequency: the leicester cough monitor," *IEEE Transactions on Biomedical Engineering*, vol. 54, no. 8, pp. 1472–1479, 2007.

[17] S. Birring, T. Fleming, S. Matos, A. Raj, D. Evans, and I. Pavord, "The leicester cough monitor: preliminary validation of an automated cough detection system in chronic cough," *European Respiratory Journal*, vol. 31, no. 5, pp. 1013–1018, 2008.

[18] J. Amoh and K. Odame, "Deepcough: A deep convolutional neural network in a wearable cough detection system," in *Biomedical Circuits and Systems Conference (BioCAS), 2015 IEEE*. IEEE, 2015, pp. 1–4.

[19] M. M. Rahman et al, "InstantRR: Instantaneous Respiratory Rate Estimation on Context-aware Mobile Devices," in *EAI International Conference on Body Area Networks*, 2018.

[20] M. Y. Ahmed et al, "mlung: Privacy-preserving naturally windowed lung activity detection for pulmonary patients," in *BSN*, 2019.

[21] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[22] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[23] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[24] L. Sicong, Z. Zimu, D. Junzhao, S. Longfei, J. Han, and X. Wang, "Ubiear: Bringing location-independent sound awareness to the hard-of-hearing people with smartphones," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 2, p. 17, 2017.