Listen2Cough: Leveraging End-to-End Deep Learning Cough Detection Model to Enhance Lung Health Assessment Using Passively Sensed Audio

XUHAI XU, University of Washington
EBRAHIM NEMATI, KOROSH VATANPARVAR, VISWAM NATHAN, TOUSIF AHMED, MD MAHBUBUR RAHMAN, DANIEL MCCAFFREY, JILONG KUANG, and JUN ALEX GAO, Digital Health Lab, Samsung Research America

The prevalence of ubiquitous computing enables new opportunities for lung health monitoring and assessment. In the past few years, there have been extensive studies on cough detection using passively sensed audio signals. However, the generalizability of a cough detection model when applied to external datasets, especially in real-world implementation, is questionable and not explored adequately. Beyond detecting coughs, researchers have looked into how cough sounds can be used in assessing lung health. However, due to the challenges in collecting both cough sounds and lung health condition ground truth, previous studies have been hindered by the limited datasets. In this paper, we propose Listen2Cough to address these gaps. We first build an end-to-end deep learning architecture using public cough sound datasets to detect coughs within raw audio recordings. We employ a pre-trained MobileNet and integrate a number of augmentation techniques to improve the generalizability of our model. Without additional fine-tuning, our model is able to achieve an F1 score of 0.948 when tested against a new clean dataset, and 0.884 on another in-the-wild noisy dataset, leading to an advantage of 5.8% and 8.4% on average over the best baseline model, respectively. Then, to mitigate the issue of limited lung health data, we propose to transform the cough detection task to lung health assessment tasks so that the rich cough data can be leveraged. Our hypothesis is that these tasks extract and utilize similar effective representation from cough sounds. We embed the cough detection model into a multi-instance learning framework with the attention mechanism and further tune the model for lung health assessment tasks. Our final model achieves an F1-score of 0.912 on healthy *v.s.* unhealthy, 0.870 on obstructive *v.s.* non-obstructive, and 0.813 on COPD *v.s.* asthma classification, outperforming the baseline by 10.7%, 6.3%, and 3.7%, respectively. Moreover, the weight value in the attention layer can be used to identify important coughs highly correlated with lung health, which can potentially provide interpretability for expert diagnosis in the future.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing**; • **Applied computing** → **Life and medical sciences**.

Additional Key Words and Phrases: Cough detection, Lung health assessment, Multi-instance learning

ACM Reference Format:
Xuhai Xu, Ebrahim Nemati, Korosh Vatanparvar, Viswam Nathan, Tousif Ahmed, Md Mahbubur Rahman, Daniel McCaffrey, Jilong Kuang, and Jun Alex Gao. 2021. Listen2Cough: Leveraging End-to-End Deep Learning Cough Detection Model to Enhance Lung Health Assessment Using Passively Sensed Audio. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 43 (March 2021), 22 pages. https://doi.org/10.1145/3448124

Authors' addresses: Xuhai Xu, xuhaixu@uw.edu, University of Washington, 1410 NE Campus Parkway, Seattle, WA; Ebrahim Nemati; Korosh Vatanparvar; Viswam Nathan; Tousif Ahmed; Md Mahbubur Rahman; Daniel McCaffrey; Jilong Kuang; Jun Alex Gao, Digital Health Lab, Samsung Research America, 665 Clyde Ave, Mountain View, CA.

2474-9567/2021/3-ART43 $15.00
https://doi.org/10.1145/3448124

1 INTRODUCTION

Ubiquitous computing is entering every part of our life. As close companions, mobile phones and wearable devices are carried by users almost anywhere at any time, thus can passively monitor and capture their daily activities and behavior. This provides an unprecedented chance for health condition assessment and monitoring. Among various physical health conditions, lung health is one of the most important topics. According to a recent report from Global Initiative for Chronic Obstructive Lung Disease (GOLD) [3], the pulmonary disease is one of the major leading causes of morbidity and mortality globally. The American Lung Association reported that over 35 million people live with at least one type of chronic lung disease across the country [4], such as chronic obstructive pulmonary disease (COPD), interstitial lung disease, asthma, tuberculosis, COVID-19, and others [42, 48]. The annual cost spent on pulmonary diseases has increased drastically in the past decade [1], and keeps growing, especially due to the recent COVID-19 pandemic situation [58].

Coughing is one of the most common symptoms of pulmonary diseases, especially for COPD and asthma patients [66, 79]. Microphones embedded in nowadays ubiquitous devices enable the sensing of cough sound. There have been rich studies about audio-based cough detection algorithms [7, 8, 13, 14, 57, 89]. However, the majority of the previous studies usually involved a specific population with limited size (*e.g.*, COPD or asthma patients), without addressing the generalizability of the cough detection model. Recently, as more cough datasets are available, FluSense [5] was able to train a deep learning cough detection model on a large amount of cough data from Google AudioSet [31] and transferred the model on their own self-collected dataset. Their model used spectrogram as the input, which is a common technique of utilizing spectral features, such as mel-filterbank or mel-frequency cepstral coefficients (MFCC). However, using spectral features for cough detection not only may hide useful information within the raw time-series signal, but also requires additional complexity on software or hardware setup. Therefore, as opposed to this method, we use an end-to-end deep learning model which takes raw audio signals as input, which simplifies both software and hardware, eliminates the assumption of the spectral content, and thus can potentially amplify the model's generalizability. By building upon the powerful pre-trained MobileNet [75] and leveraging various augmentation techniques, our method can achieve significantly better performance without any fine-tuning (5.8% on a clean dataset and 8.4% on a noisy in-the-wild dataset).

Detecting cough is the first step towards passive lung health assessment, *i.e.*, prediction of a subject's certain lung health condition category (*e.g.*, obstructive pulmonary disease *v.s.* non-obstructive pulmonary disease). The cough sounds contain rich information that reflects the condition of the lung, such as the cough frequency and energy, whether there is sputum or mucus in the lung airways, *etc.* Recently, researchers have started to investigate how to use cough sounds for various lung health assessment tasks, such as lung health diagnosis and pulmonary disease type classification. However, due to the difficulties of collecting both cough sounds and ground truth labels, previous studies usually had a limited dataset and could only resort to traditional machine learning models. To mitigate this issue, we observe the richness of the cough data, and thus propose a new method to transfer the feature representation of cough detection task to the lung health assessment tasks so that existing large-scale cough sound data can be effectively exploited. Our hypothesis is that all of these tasks need effective feature embedding from the deep learning model. Therefore, the model trained for cough detection can potentially be easily adapted for assessment tasks, and the cough detection model's feature embedding can be an effective initialization for the assessment models. As one user can have multiple cough sounds, an attempt to simply take an average of them will lead to information loss. To address this issue, we employ a multi-instance learning (MIL) framework with the attention mechanism and build new end-to-end deep learning models for the assessment tasks. To the best of our knowledge, we are the first to propose to connect the cough detection and lung health assessment tasks. In this paper, we focus on three classification tasks: healthy *v.s.* unhealthy, obstructive *v.s.* non-obstructive, and COPD *v.s.* asthma classifications. Our models achieve an F1-score of 0.912, 0.870, and 0.813 on the three tasks, respectively. Compared to the best baseline, our method outperforms by

10.8%, 6.3%, and 3.7%, respectively. In addition, the attention weights in the MIL framework imply the relative importance of a user's cough sounds for each task, which can provide interpretability and identify a small set of cough sounds that may help experts for further diagnosis.

Our main contributions in this paper are three-fold:

- We built an end-to-end audio-based cough detection model by leveraging pre-trained MobileNet and various audio augmentation techniques. Our model is able to achieve an F1 score of 0.948 on a clean dataset and 0.884 on a noisy dataset without any fine-tuning, outperforming the baseline model by 5.8% on the first dataset and 8.4% on the second dataset.
- Based on the detection model, we are the first to propose to transfer the cough detection task to lung health assessment tasks. Using a MIL model with the attention mechanism, our new models outperform baselines by 10.8%, 6.3%, and 3.7% on healthy *v.s.* unhealthy, obstructive *v.s.* non-obstructive, and COPD *v.s.* asthma classification tasks, respectively.
- We further leverage the attention weights in the MIL framework to identify the important cough sounds highly correlated with lung health condition for the assessment tasks, which can provide further interpretability during diagnosis.

2 BACKGROUND

In this section, we first review the recent advances on audio-based cough detection methods (Section 2.1) and pulmonary-related disease diagnosis via mobile and wearable technologies (Section 2.2). We then briefly introduce MIL, attention mechanism, and their applications (Section 2.3).

2.1 Cough Detection from Audio Signals

Cough is one of the most prominent symptoms associated with many respiratory diseases such as COPD, asthma, tuberculosis, gastro-oesophageal reflux, cystic fibrosis, and chronic bronchitis [11, 61]. Since the 1950s, there has been extensive research on cough detection using audio signals [13]. Researchers have proposed a large number of algorithms for automatic cough detection, such as support vector machine (SVM) [30], hidden Markov models (HMM) [14], random forest (RF) [50], k-nearest neighbor (kNN) [59], Gaussian mixture models (GMM) [54], time delay neural network (TDNN) [8], convolutional neural network (CNN) [6], and recurrent neural network (RNN) [7]. Most of the previous works rely on feature engineering. Specific acoustic features need to be extracted from raw audio signals for these cough detection algorithms. For example, Matos *et al.* [57] and Zhu *et al.* [89] extracted MFCC features to train HMM models. Larson *et al.* [50] build RF models using spectrogram-based features and Amoh and Odame [7] ran deep learning models on spectrogram images. Besides, other spectral features such as Hilbert marginal spectrum [51] and Gammatone cepstral coefficient [53] have also been explored. There are several limitations in previous works. For example, most datasets used in those studies were collected from a limited number of participants in a specific population (*e.g.*, COPD or asthma patients), without evaluating the generalizability of a cough detection model, *i.e.*, training models from one dataset and testing them on a completely different dataset. Recently, more cough datasets are emerging and some can be expected to become publicly available soon [2, 77], especially due to the COVID-19 pandemics. Perhaps the closest existing work is FluSense [5]. They trained a CNN model on the large-scale cough data in Google AudioSet [31] and tested it on a newly collected in-the-wild dataset. Their model achieved an F1 score of 0.750 when the model was applied directly. Although involving large-scale datasets, their model still requires further fine-tuning on the new dataset to achieve an F1 score of 0.880.

In contrast to the common technique of using spectral features such as mel-filterbank or MFCC (also adopted by FluSense), we focus on training an end-to-end model from multiple public datasets [31]. Although end-to-end model that takes raw time-domain input waveform as the input is commonly used for tasks such as scene

classification and auditory object recognition [9], there has been very few studies specifically focusing on building end-to-end cough detection models [25]. An end-to-end model not only simplifies the software or hardware complexity as it does not require to calculate and extract any spectral features, but more importantly, it obviates the assumption of spectral content so that features can be optimized for tasks, which can help improve the generalizability of a model [28, 39]. Our cough detection model achieved an F1 score of 0.884 without any fine-tuning when applied on a new noisy in-the-wild dataset.

2.2 Pulmonary Condition Assessment Using Ubiquitous Devices

With the advances of ubiquitous computing in the past decade, researchers started to investigate pulmonary disease diagnosis using mobile phones or wearable devices [18, 19, 21, 40, 44, 60, 85]. Multiple sensing modalities have been explored. For example, Rahman *et al.* [71] leveraged HR/HRV data from chest-bands and built models to diagnose pulmonary patients. Juen *et al.* [45] proposed to capture patients' gait speed in a 6-min walk test with an accelerometer for COPD diagnosis. Because of the fact that audio is one of the most widely available modalities on both stationary and mobile devices, and that cough is a common indicator of many pulmonary diseases, there have been some recent studies that used cough sounds for lung health assessment [34, 35, 42, 72, 74, 90]. For example, Pramono *et al.* [67] proposed to use acoustic features of cough to detect pertussis with an accuracy of 92.0% on 38 patients. Laguarta *et al.* [48] recently showed that cough sounds can be used for COVID-19 detection with an AUC of 0.97. Sharan *et al.* [76] conducted severity classification for COPD patients using cough sounds with an accuracy of 70.0%. Similarly, Nemati *et al.* [61] used cough sounds to perform disease detection and severity classification for COPD and achieved 91.0% and 95.0% accuracy, respectively[1].

However, due to the difficulties of collecting both lung health ground truth labels and cough sounds from the same person, prior studies had limited data size. Therefore, they usually resorted to feature engineering and trained traditional off-the-shelf machine learning models from scratch using these features (*e.g.*, SVM and RF) [61, 76]. Meanwhile, as stated in Section 2.1, we found that large-scale cough sound data can be expected. As both the cough detection task and lung health assessment tasks share similarities in terms of the feature embedding, the model trained for cough detection has the potential to be tuned for assessment tasks. Therefore, we propose to transfer from the cough detection task to lung heath assessment tasks so that rich cough data can be leveraged to address the lack of data issues. As we show in Section 5, our model significantly outperforms the baseline, indicating the effectiveness of our method.

2.3 Multi-Instance Learning and Attention Mechanism

Multiple instance learning (MIL) is a form of weakly supervised learning [27, 87]. It deals with data arranged in sets (usually named bags). Labels are only provided for entire sets, and the specific labels of the instances in the bags are not available [16]. The goal is usually to predict the bag labels. Such a formulation is suitable for a wide range of areas, such as drug activity prediction [12], document classification [64], web mining [88], and object localization in images or videos [24]. More related to our paper, MIL is also often used for computer-aided diagnosis. Bag labels, such as the overall diagnosis of a patient, are usually easier to obtain than instance labels, such as outlines of abnormalities in a medical scan [69, 82, 84]. For example, Cheplygina *et al.* [22] used a MIL classifier trained on chest computed tomography images for the task of classification of COPD. There are relatively less studies using MIL on audio signals [17, 73]. In our work, MIL naturally fits in the lung health assessment tasks, where a subject's lung health status is the bag label and their multiple sound snippets are the instances within the bag. We used deep MIL, where each sound snippet is converted to a vector embedding and multiple sounds of one subject aggregated to produce the outcome. Our experiment shows that compared to an instance-based model, such a framework greatly improves the classification result.

[1]Note that these accuracy results are not directly comparable due to the difference between datasets

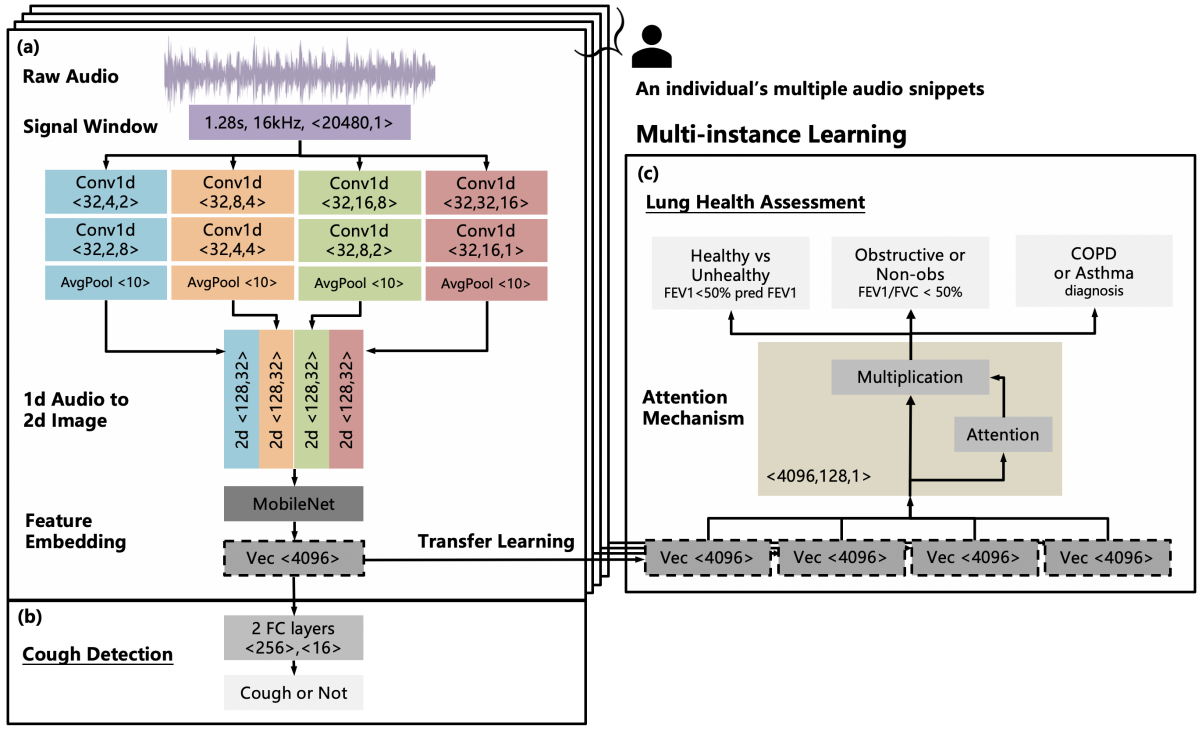


Fig. 1. The overview of the framework of Listen2Cough. (a) The end-to-end deep learning architecture to extract feature representation from raw audio recordings. (b) The cough detection part. Combining (a) and (b) leads to the cough detection model. (c) The lung health assessment part in MIL framework with attention mechanism. Combining (a) and (c) leads to the lung health assessment models that are transferred from the cough detection model. For each conv1d, the hyperparameters are shown as <channels,size,strides>. The FC's hidden size is shown as <size>. And the attention layer's hyperparameters are shown as <input size,hidden size,output size>.

Attention mechanism in deep learning can be abstracted as a vector of importance weights [10]: to predict or infer one element (*e.g.*, a pixel in an image or a word in a sentence), a model uses the attention vector to determine how strongly it is correlated with (usually referred as "attends to") other elements and take the weighted-sum as the target approximation [70]. It can not only improve the prediction or inference performance, but also provide interpretability as its attention weights indicate which part of the input gets higher importance for a task [20, 23, 52]. We integrated the attention mechanism into the MIL framework [41] for lung health assessment tasks. Our analysis showed that the attention weights can help identify important variation of cough sounds for different assessment tasks, which can potentially assist medical experts for further diagnosis.

3 LISTEN2COUGH METHOD

In this section, we introduce our method for both the cough detection and lung health assessment tasks in detail. Figure 1 shows the overall framework. We first present the method of building generalizable cough detection models (Section 3.1). Then, we describe how we transfer from the cough detection task to the lung health assessment tasks (Section 3.2).

3.1 Generalizable Cough Detection Model

We combine two efforts to build a cough detection model that is robust to various noise types, consistent across multiple devices, and generalizable to new external datasets: 1) an end-to-end architecture that combines 1d-CNN and MobileNet [75] (Section 3.1.1), and 2) multiple augmentation techniques to further enhance the model robustness (Section 3.1.2).

3.1.1 End-to-end Model Architecture. As mentioned in Section 2.1, there are several advantages of an end-to-end model. Taking the raw audio signal as the input, we choose a window size of 1.28 seconds (mono channel audio sampled at 16kHz, thus the input is a 1d vector with length of 20480) to cover majority of cough samples [61].

Our model starts with two 1d-CNN layers to extract effective temporal and spectral features. A cough typically involves three steps: 1) an initial deep inspiration and glottal closure, 2) the contraction of the expiratory muscles, and 3) a sudden glottis opening with an explosive expiration. However, it remains unclear that during this procedure, what temporal scope (reflected in the kernel size) a convolutional layer should have for effective detection. To address this issue, we adopt an Inception-like nucleus [80] with four parts of convolution layers, each with different kernel sizes so that they can focus on features at different temporal resolutions [28]. We choose 4,8,16,32 for the first layer, and half for the second layer (2,4,8,16); these are common numbers used in existing audio-based deep learning studies (*e.g.*, [9, 25, 39]). The stride of the first layer is set as half of the kernel and the stride of the second layer is set accordingly such that the output of these four parts will have the same length. Then, these four parts are concatenated together to form a 128x128 2d image.

We feed the image to a MobileNet [75] that is pre-trained on ImageNet dataset [26, 47]. Previous studies have shown that the model trained on ImageNet can generalize well across datasets [46]. We choose MobileNet since it is well known for its reduction of the parameter number compared to counterparts such as VGG [78] or ResNet [37], thus suitable for mobile computing. The parameters of the MobileNet are trainable by backpropagation so that it can be further optimized for the cough detection task. We only retain the embedding of the MobileNet's final convolution layers (with the size of 4096), discard the rest of the fully connected (FC) layers, and concatenate two new FC layers (with the hidden size of 256 and 16) at the end to generate cough detection outcomes. ReLU [33] is used as the activation function in the model. We employ binary cross-entropy as the loss function and set all parameters in the model to be trainable. Figure 1(a) and (b) presents the final architecture together with the hyperparameters.

3.1.2 Augmentation Techniques. In addition to building an effective end-to-end deep learning architecture, we also leverage a number of augmentation techniques to further enhance the model's generalizability. Note that instead of augmenting the data beforehand to enlarge the dataset size (*e.g.*, [5, 49]), we apply these techniques on the raw audio of each input batch *during* the training. These can increase the variance of the data in the training process, producing a more robust and generalizable model.

Re-sampling: As we will introduce in Section 4, although the dataset has a large amount of cough data, it is still quite unbalanced. Therefore, we intentionally over-sample cough sounds during the training. Moreover, we also over-sample noises that are close to cough sound (*e.g.*, dog bark, door close, sneeze). Specifically, for each input batch, we over-sample cough and cough-like noise such that the relative proportion of cough:cough-like noise:other noise is 2:1:2. In such a way, the model can be trained on a more balanced but high-variation set, thus enhancing the classification ability of the model.

Standard audio adjustment: We apply basic standard adjustment on the raw audio signal before feeding it into the model [55, 63], including amplitude shift (x0.25 to x4, to simulate audio events at different difference), time shift (-0.5s to +0.5s rolling basis, to simulate different signal window positions in the audio stream), as well as pitch shift and speed shift (both x0.5 to x2, to simulate multiple variations of audio events). The corresponding adjustment parameters are determined by a random number uniformly distributed within the ranges. Moreover,

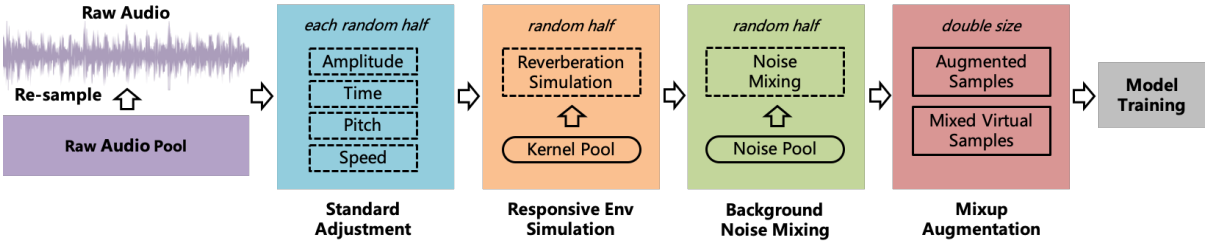


Fig. 2. The audio augmentation pipeline. A box with dashed borders indicate that this augmentation technique is applied on a random half of the input audio batch.

for each of the four adjustments, we randomly pick half of the input batch and apply it to the data. Therefore, within each batch, there are approximately 1/16 (2^4) of the audio input do not get any adjustment, 1/16 only get one of the four shifts... and so on. Such a method maximizes the variance of the input data during the training.

Responsive environment simulation: Different environments have different acoustic properties (*e.g.*, classroom, hallway, auditorium), which can greatly impact the collected audio. Therefore, we also add reverberation to the raw audio to simulate the sound in multiple responsive environments. To simulate the situation, each time we randomly pick a kernel sample from Aachen room impulse response (AIR) dataset [43] and then calculate the convolution with the raw audio. We randomly apply this technique on half of the input batch. Note that this technique is applied independently of the standard augmentation technique.

Background noise mixing: Besides the augmentation on the audio itself, we also apply the common background noise mixing technique on the audio to ensure the model's robustness towards different noisy environments (*e.g.*, kitchen, street, station). Each time we randomly pick a slice at the same length (1.28s) from a noise audio pool, and mix it with the raw audio at a power rate uniformly distributed between 0 to 1 (0 as no noise and 1 as same power between the noise and the raw audio). Similarly, we also apply this technique to the random half of the input batch, independent of the above two techniques.

Mixup: Mixup is a recent augmentation technique to improve a model's generalization by diversifying the training distribution [86]. It constructs virtual training samples by a pair of samples in the original training set. Specifically, given two samples, (x_i, y_i) and (x_j, y_j), the virtual sample $(\widetilde{x}, \widetilde{y})$ is determined by

$$\widetilde{x} = \lambda x_i + (1 - \lambda) x_j$$

$$\widetilde{y} = \lambda y_i + (1 - \lambda) y_j$$

where $\lambda \sim beta(\alpha, \alpha)$. We set α to be 1 so that the $beta$ distribution degrades to a uniform distribution to ensure unbiased mixup. For example, if two cough snippets are mixed, the virtual sample would sound like two person coughing at the same time and the virtual label is still positive. While if one cough and one noise snippet are mixed, it would then sound like a cough in a noisy background, with the label being half positive and half negative. Given an input batch, each time we randomly pick two audio windows and generate a virtual sample. We repeat this multiple times until the number of virtual samples equals the batch size, *i.e.*, the final input batch is doubled by the mixup technique.

These augmentation techniques are combined together to amplify the robustness and the generalizability of the cough detection model. Figure 2 visualizes the procedure

3.2 Connecting Cough Detection and Lung Health Assessment

Detecting cough is the prerequisite of passive lung health monitoring. The next step is to leverage cough sounds to assess lung health conditions. Previous work has shown that cough audio is meaningful for lung health

assessment [61, 76] but their models are limited by the small size of the dataset. To address this problem, we propose to transfer the model parameters from the cough detection task to leverage rich cough data. We first briefly describe the assessment tasks in Section 3.2.1. Then, we introduce our transferring method in detail.

3.2.1 Lung Health Assessment Tasks. To evaluate the functionality and condition of a lung, common respiratory parameters are used such as forced vital capacity (FVC), forced expiratory volume in the first second (FEV1), and FEV1/FVC ratio which represents the obstruction level of the respiratory airways in the lung. Therefore, we focus on three lung health assessment tasks in our study:

- **Healthy *v.s.* Unhealthy**: We first classify the lung health condition based on FEV1; we refer to the Global Initiative for Lung Disease (GOLD) criteria [3]: healthy (FEV1 $\geq$ 50% predicted FEV1) and unhealthy (FEV1 $<$ 50% predicted FEV1).
- **Obstructive *v.s.* Non-obstructive**: The second task is to determine whether the respiratory disease is obstructive (FEV1/FVC < 0.7) or not (FEV1/FVC ≥ 0.7). This is an important diagnosis decision often made by doctors that would be used in adjusting the patients' treatment [56].
- **COPD *v.s.* Asthma**: In obstructive lung diseases, the most common two types are COPD and asthma. Therefore, the third task is to distinguish COPD patients and asthma patients.

3.2.2 From Cough Detection to Lung Health Assessment. In prior studies regarding lung health assessment, researchers extracted and utilized both time-domain and frequency-domain features (*e.g.*, zero crossing rate, range of the amplitude, spectrogram, MFCC). Due to the limited data size, they trained traditional machine learning classifiers for these tasks [61, 76]. Observing the rich cough data for cough detection, as well as the flexible deep learning architecture introduced in Section 3.1.1, the models for the three lung health assessment tasks can be transferred from the end-to-end cough detection model.

As one subject usually has more than one cough sound snippets, one subject can be regarded as a bag, and the multiple cough sounds belonging to this subject are instances in the bag, which makes the MIL framework a perfect fit. Therefore, for one subject's all cough data, we feed each sound snippet into the cough detection model, take the feature embedding from the output of the MobileNet part, and aggregate them with a deep MIL framework. We use an attention mechanism for the deep MIL framework [41] for effective aggregation and interpretability. Specifically, let $V = \{\mathbf{v}_1, \mathbf{v}_2, ...\mathbf{v}_N\}$ be a bag of N instances, with each V_i being the k-dimension embedding of a cough audio window. The MIL pooling is

$$\mathbf{z} = \sum_{n=1}^{N} \alpha_n \mathbf{v}_n$$

$$\alpha_n = \frac{\exp\{\mathbf{w}^T \tanh(\boldsymbol{M}\mathbf{v}_n)\}}{\sum_{i=1}^{N} \exp\{\mathbf{w}^T \tanh(\boldsymbol{M}\mathbf{v}_i)\}}$$

where $\mathbf{w} \in \mathbb{R}^{L\times 1}$ and $\boldsymbol{M} \in \mathbb{R}^{L\times k}$ are parameters to be trained. Following Ilse *et al.* [41], we also employ the hyperbolic tangent $\tanh(\cdot)$ to include both negative and positive values for proper gradient flow, and set the hidden attention dimension L as 128 and the output dimension as 1. For each lung health assessment task, we use binary cross-entropy as the loss function. All parameters, including the 1d-convolutional layers, the MobileNet, and attention weights, are trainable and can be further optimized for the assessment tasks. In other words, we used the cough detection model's parameters as meaningful initialization for the new models. The right part of Figure 1 visualizes the framework.

Table 1. Public datasets for building robust and generalizable cough detection model.

Name	Purpose	Contents	Duration (seconds)
AudioSet (labelled) [5]	Cough Data, Negative Samples, Re-sampling	Cough, sneeze, speech, sniffle, and other noise	45550
ESC-50 [65]	Cough Data, Negative Samples, Re-sampling	Cough, breath, snore, laugh, and other noises	10000
Freesound [29]	Cough Data, Negative Samples, Re-sampling	Cough, sneeze, speech, laugh, and other noises	7020
ESTI [68]	Negative Samples, Noise Mixing	Background Noise	1400
DEMAND [81]	Negative Samples, Noise Mixing	Background Noise	86400
AIR [43]	Responsive Environment Simulation	Convolution Kernel	-

4 DATA COLLECTION AND IMPLEMENTATION

We first introduce several public datasets we used for model training (Section 4.1) and the two datasets we collected to demonstrate our method's effectiveness and generalizability (Section 4.2). Then, we briefly describe the implementation of our methods (Section 4.3).

4.1 Public Datasets

Two types of datasets were involved in building an effective cough detection model: cough datasets and augmentation datasets. For cough datasets, we leveraged the labeled Google AudioSet [31] (FluSense [5] generously publicized their manual labels). Moreover, ESC-50 [65] and Freesound [29] also contained some cough data. We use these data as cough positive samples. Meanwhile, other noise sounds in these datasets were used as negative samples. As mentioned in Section 3.1.2, we manually labeled some noise types in these datasets as cough-like negative samples. Together with cough sounds, these noise signals will be over-sampled during the training.

For augmentation datasets, we employed ESTI binaural background noise database [68] and Diverse Environments Multichannel Acoustic Noise Database (DEMAND) [81] as the background noise datasets. We used one half as negative samples (not cough-like), and the other half as the noise mixing pool. Moreover, AIR [43] was employed as the convolution kernel pool for reverberation augmentation. Table 1 summarizes the purpose and the contents of all datasets.

4.2 Multiple Data Collection Studies

In addition to the public datasets, we also conducted two in-clinic data collection studies in collaboration with our partner pulmonologists from hospitals.

4.2.1 Participants. The first study is a controlled clinical study during patients' regular visits in a quiet environment. We recruited 131 participants (64 female, 67 male, age 43±19). The second study was conducted during uncontrolled clinical visits and contained a large amount of natural noise. Another group of 70 participants were recruited (41 female, 29 male, age 55±16).

Participants with COPD needed a physician diagnosis of COPD. Similarly, participants with asthma needed to have a physician diagnosis of asthma. In our studies, asthmatic participants could not have a smoking history greater than 20 pack-years to avoid the bias introduced by its effect on vocal cords. Participants were excluded due to the safety concern if they had a history of congestive heart failure with New York Heart Association (NYHA) Classification II or greater symptoms, had a history of pneumonectomy, vocal cord dysfunction, or laryngeal surgery, or were pregnant. In the first study, 40 did not have pulmonary diseases and 91 were chronic pulmonary patients. 69 of them were diagnosed with asthma, 9 with COPD, and 13 exhibited co-morbidity of both asthma and COPD. In the second study, 25 participants had COPD. 25 had asthma. 10 had chronic cough, and 10 did not have pulmonary-related symptoms. Table 2 summarizes the demographics of the two datasets.

Table 2. Demographic information and health conditions of the two datasets. Due to the lack of valid pulmonary parameters measurement, some of the information from dataset1 participants' is not available. Both datasets are used to evaluate our cough detection model, while only dataset2 is used for the lung health assessment tasks.

	# of Subjects	Healthy *v.s.* Unhealthy	COPD *v.s.* Asthma	Obstructive *v.s.* Non-obstructive
Dataset1	131	-	-	-
Dataset2	70	46 *v.s.* 24	25 *v.s.* 25	43 *v.s.* 27
	Age	Gender	Height	Weight
Dataset1	43±19	F 64, M 67	168±15 cm	83±25 kg
Dataset2	55±16	F 41, M 29	166±18 cm	85±29 kg

4.2.2 Procedure. Both studies were conducted in a hospital with the supervision of the pulmonologists and were approved by the Institutional Review Board (IRB). They followed a similar procedure. During each study, participants went through a series of tasks, such as tidal breathing (1-min), supine breathing (1-min), scripted speech or reading (a given paragraph of text, 1-min), a spontaneous speech session (1-min), and a voluntary cough session (2-min). Participants were offered a Samsung Note 8 smartphone. The smartphone was held in hand and collected audio data via the microphone.

There were also some differences between the two studies. In the first study, a research assistant (RA) was also present and used another smartphone to label the start and the end of each task. The microphone in the RA's phone was also turned on thorough out the study. Therefore, we collected audio from two separate devices in the first study. In the second study, we collected valid ground truth of participants' lung function. Participants went through a spirometry test using a hospital-grade spirometer (Pneumotrac Portable Screening Spirometer-model 6800) where lung function values including but not limited to FEV1, FVC, and FEV1/FVC were collected. With their demographics, FEV1% and FVC% are also calculated. Thus the lung health ground truth labels were available for the second study. Comparatively, we lacked valid pulmonary parameters in the first study. Thus, we left a few blank cells for dataset1 in Table 2. Therefore, both datasets were used for cough detection verification and only the second dataset was used for lung health assessment tasks.

4.2.3 Annotation. In order to minimize the error in cough segmentation, data from cough sessions were manually annotated and segmented by trained annotators after the data collection. The Audacity toolbox was utilized which enabled the annotators to simultaneously listen to and visualize the cough sounds within the audio recordings in order to tag the start and end times of each cough.

4.3 Implementation of Our Method

Having both public datasets and self-collected datasets, we applied the same pre-processing steps on the data. All audio signals were normalized between -1 to 1 floating-point value at a 16k Hz sampling rate. The signal window was set as 1.28 s (1.28×16000 = 20480), which was long enough to cover the majority of cough samples. For an audio snippet shorter than 1.28 s, the signal window was centered at the middle of the snippet. For a snippet that was longer than 1.28 s, *e.g.*, a long cough sound, the signal window moved with the hop size as 0.64 s until the end of the cough sound, generating multiple samples from the snippet. Empty parts were padded with zero (at the beginning or the end of the audio). For cough detection, prior to any augmentation, the number of original positive cough samples in public datasets is about 14k, and the number of negative cough samples is about 100k. We used 90% of them as the training set and 10% of them as the validation set. Our testing set is the data collected from the two user studies, which contains around 6/8k positive cough samples and 30/35k negative samples in

the first/second study. For lung health assessment, each participants has 120±54 cough windows on average collected (the second study only). We employed a five-fold cross-validation to reduce variance of the results.

We used Keras with Tensorflow backend [32] to implement all models, and trained the models using four Nvidia Tesla K80 cores (11G RAM). For the cough detection model, all cough windows were marked as positive samples and others were marked as negative samples, with binary cross-entropy as the loss function. We set the original batch size as 32. After the augmentation, the size for each batch became 64. Meanwhile, all cough sounds were linked with the speaker. For the lung health assessment models, we set the batch size as 1 (one subject per batch) and all cough sounds belonging to this subject were included. The loss functions of the three tasks were all bag-level binary cross-entropy. For all models, we used an Adam optimizer with a dynamic learning rate schedule that started from 0.001 and repeatedly reduced to half once the validation accuracy did not improve after 10 epochs. The epoch number was set as 200, with an early stop number as 50.

5 RESULTS

In this section, we first present the results from the cough detection model on both self-collected datasets in Section 5.1. Then, using the second dataset, we summarize the results of the transferred models on the three lung health assessment tasks in Section 5.2.

5.1 Cough Detection

We compared the performance of our model with the following two baselines:

1) **Feature Engineering**: Prior to the deep learning era, researchers often resorted to the typical signal processing method, *i.e.*, feature engineering, and then trained off-the-shelf machine learning models for detection. Sometimes this method can achieve good results, especially when the data amount is limited and the sound has little background noise. Following a recent work [62], we extracted a number of time-domain features and frequency-domain features within the 1.28 s window, such as total energy, zero crossing, MFCC features, spectral variance, kurtosis, and skewness, *etc.* Then, we trained an RF classifier (100 trees) using these features.

2) **Spectrogram and CNN**: One of the most common methods in recent cough detection studies tends to leverage CNN models on the spectrogram. We chose a recent work FluSense [5] as the baseline since they also investigated the generalizability of the cough detection model. We picked the model having the best performance in their paper and trained using our datasets.

In the rest of this section, we evaluated our model and baselines by answering a series of questions.

5.1.1 How Does the Model Perform on Clean Data? We first tested our method on a clean dataset that contained little noise. We used the public cough data (Section 4.1) as the training set, and the first study's data (Section 4.2) as the testing set, including the audio collected from both the subject's and the RA's phones. Table 3 summarizes the results.

Our model shows better performance than the CNN model, followed by the traditional ML model. Compared to the CNN baseline, our method achieved close recall (only 0.5%) but much higher precision (11.4%), leading to an advantage of 5.8% on F1 score and 5.1% on balanced accuracy. This indicates that our method is more robust to noises and produces less false positive samples.

5.1.2 How Does the Model Perform Differently between Devices? As the first dataset contains audio collected from two devices, we also tested them separately, *i.e.*, data from each device was used as the testing set, to evaluate our model's robustness towards device variation. Table 4 presents the results. Generally, our model achieved good results on both devices, with small distinction (1.3% on F1 score). This validated the robustness of our method. Close inspection of the performance also revealed some interesting findings. When applying our method on the data from the subject's phone, the result showed slightly higher recall but lower precision. This might be

Table 3. Cough detection results on the first dataset, where the audio is collected in a quiet environment without much noise. Note that here we include the data collected by both the phone in the subject's hand and the phone the RA's hand.

Setup	Bal Acc	Prec	Rec	F1
Feature Engineering [62]	0.879	0.861	0.846	0.853
Spectrogram and CNN [5]	0.895	0.873	0.907	0.890
Our end-to-end model	0.949	0.987	0.912	0.948
delta	5.1%	11.4%	0.5%	5.8%

explained by the fact that the subject's phone was closer to the cough source (the subject themselves), which could make the cough sound clearer and the cough detection easier, leading to a higher recall. In contrast, the RA's phone was closer to multiple noise sources (*e.g.*, the desk, the door, the hallway). Although noises were not very common in this dataset, clearer noise sounds could simplify the noise detection, thus resulting in higher precision.

Table 4. Cough detection results on two devices' data in the first dataset separately.

Setup	Bal Acc	Prec	Rec	F1
Data from the subject's phone	0.954	0.959	0.955	0.957
Data from the RA's phone	0.932	0.986	0.906	0.944

5.1.3 How Does the Model Perform on Noisy Data? We evaluated the performance of our model on a noisy dataset that is closer to real-life scenarios. After training the model using the public cough datasets, we then tested it on our second dataset where a large number of natural noises were included. Table 5 summarizes the results.

Unsurprisingly, all models' performance decreased due to the lower signal-to-noise ratio (SNR) on the noisy audio. But similar to the results on the first dataset, our model consistently outperformed the baselines. Compared to the CNN baseline, our model had a slightly better performance on recall (3.4%) and much better performance on precision (13.7%), leading to better results on F1 score by 8.4% and balanced accuracy by 3.7%. The advantage of our method was bigger on the noisy dataset than on the clean dataset (see Table 3). This enhances our findings in Section 5.1.1 that our model is more robust to various noise types compared to previous methods.

Table 5. Cough detection results on the second dataset, where the audio is collected in a noisy clinical environment with a large number of natural noises.

Setup	Bal Acc	Prec	Rec	F1
Feature Engineering [62]	0.702	0.738	0.719	0.728
Spectrogram and CNN [5]	0.788	0.796	0.804	0.800
Our end-to-end model	0.825	0.936	0.838	0.884
delta	3.7%	13.7%	3.4%	8.4%

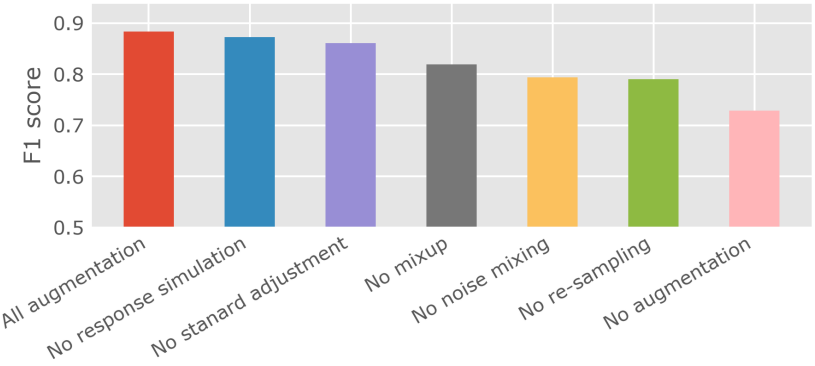


Fig. 3. The results of different augmentation technique combinations.

5.1.4 How Does Each Augmentation Improve the Model? Another interesting aspect is to investigate how each of the augmentation techniques helps improve the performance of our model. We conducted an augmentation ablation study, *i.e.*, each time we removed one of the five augmentation techniques listed in Section 3.1.2 and re-trained the model: re-sampling, standard adjustment, response simulation, noise mixing, and mixup. We also trained an additional model without any augmentation technique as a reference model. Then, we tested the model on the second dataset. Figure 3 shows the F1 score performance of these models.

The re-sampling and noise mixing appeared to be the most effective augmentation techniques. Removing either of them would lead to a great drop in the performance (9.4% for the re-sampling and 9.0% for the noise mixing). In contrast, the response simulation had the least impact on the final performance. Removing it would lead to a slight decrease in F1 score (1.1%). The other two techniques were in the middle. Not applying the standard adjustment or the mixup would result in a drop on F1 score by 2.3% and 6.5%, respectively.

5.2 Lung Health Assessment

More importantly, we further evaluated our method's performance on the three lung health assessment tasks. After training the cough detection model on the public datasets, we connected it with the MIL framework and applied transfer learning for the assessment tasks, *i.e.*, we used the cough detection model's parameters to initialize the MIL framework and further trained the model using our dataset (collected from the second user study, where we collected valid ground truth labels). Compared to the cough detection model, the assessment models need less training time and also converged faster.

5.2.1 How Does Our Model Perform Compared to the Method Adopted by Prior Work? As previous works mainly involved only a small number of patients from a specific population, they usually resorted to the signal processing method that is similar to the first baseline described in Section 5.1: a number of features were extracted from each cough sound and an instance-based model was trained for the classification tasks [61, 76]. In our work, we also adopted this method as our baseline. Specifically, following [61], we extracted both time-domain features (*e.g.*, cough duration, zero-crossing rate, interquartile range of the amplitude) and frequency-domain features (*e.g.*, spectrogram spread, spectrogram centroid, spectrogram flatness, spectrogram rolloff, 20 MFCCs, 12 Chroma). We then trained three RF classifiers also using five-fold cross-validation, with one for each task. The results of the three tasks are shown in Table 6.

Overall, our model consistently outperformed the baseline in all three tasks, with an advantage of 10.7% (health *v.s.* unhealthy), 6.3% (obstructive *v.s.* non-obstructive), and 3.7% (COPD *v.s.* Asthma) on F1 score, respectively. This

Table 6. Three lung health assessment tasks results on the second dataset.

Setup	Task	Bal Acc	Prec	Rec	F1
Feature Engineering	Healthy *v.s.* Unhealthy	0.792	0.792	0.819	0.805
Transferred MIL	Healthy *v.s.* Unhealthy	0.815	0.867	0.963	0.912
	delta	2.3%	7.5%	14.4%	10.7%
Feature Engineering	Obstructive *v.s.* Non-obstructive	0.792	0.811	0.803	0.807
Transferred MIL	Obstructive *v.s.* Non-obstructive	0.801	0.909	0.833	0.870
	delta	0.9%	9.8%	3.0%	6.3%
Feature Engineering	COPD *v.s.* Asthma	0.751	0.764	0.788	0.776
Transferred MIL	COPD *v.s.* Asthma	0.790	0.765	0.867	0.813
	delta	3.9%	0.1%	7.9%	3.7%

indicates that our method of transfer learning and MIL framework works effectively for lung health assessment tasks. However, it remains unclear about the contribution of the two parts separately as they are integrated together in our current model, leading to the next question.

5.2.2 How Does Each Component of Our Method Contribute to the Model? Our proposed method contains two parts that complemented each other: 1) transfer learning that leverages the cough detection model trained on rich cough data, and 2) MIL framework with attention mechanism that is suitable for the setup of the assessment tasks. To evaluate their contribution separately, we compared all possible four combinations of the two components: an instance-based deep learning architecture (the left side of Figure 1, similar to the cough detection model) trained from scratch, the same architecture but transferred from cough detection (transfer learning only), the MIL framework trained from scratch, as well as the complete model in Section 5.2.1. We also conducted a five-fold cross-validation for these four models on each assessment task. Figure 4 summarizes the F1 score results of all models.

There are a few observations. First, we found that in all three tasks, no matter using the MIL framework or the instance-based CNN architecture, the transfer learning technique constantly improved the performance. In the instance-based models, leveraging the cough detection model improved the F1 score by 6.6% on the healthy *v.s.* unhealthy recognition, 1.3% on the obstructive *v.s.* non-obstructive detection, and 4.6% on the COPD *v.s.* Asthma classification. The improvement became even larger when using the MIL framework, with an advantage of 9.1%, 4.3%, and 7.3%, respectively. This shows that the parameters in the cough detection model can be a good initialization for the assessment models. Moreover, it validates our intuition that the feature embedding trained for the cough detection task is meaningful for the lung assessment tasks as well.

Second, the MIL framework also greatly improved the results. When not using transfer learning, the MIL led to the advantage of 13.9%, 11.8%, and 11.2% on the three tasks. The improvements were slightly better when transfer learning was used, with the advantage of 16.3%, 14.8%, and 13.9%, respectively. Even without any transfer, such a framework still outperformed the corresponding instance-based CNN model built on the cough detection model (7.2%, 10.5%, and 6.6%). This showed the appropriateness and the effectiveness of using MIL for these types of tasks. When combining these two components together in our method, both parts contributes to the final good performance of our models.

Moreover, to investigate the contribution of attention mechanism specifically, we further trained MIL-based models without attention mechanism (also transferred from cough detection). Compared to the complete models, the ones without attention had slightly lower performance, with a decrease of 1.5%, 2.4%, and 1.8% on the three tasks. Although the improvement is minimal, this indicates that the attention mechanism also help increase the

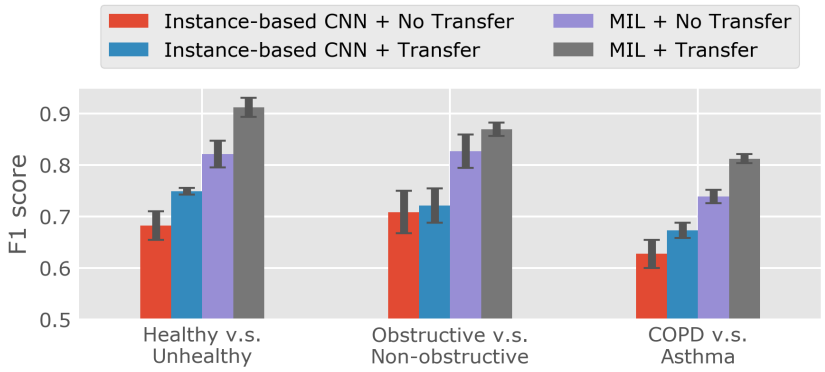


Fig. 4. The lung health assessment results of different combinations of MIL and transfer learning. Error bar indicates the standard error of the F1 score in the five-fold cross validation.

capability of the model a little. More importantly, the weights in the attention layer can provide more insights of these assessment tasks, as detailed in Section 5.2.3.

5.2.3 How Can Our Method Provide Interpretability? In addition to achieving better performance, the attention mechanism in our MIL framework can further provide interpretability to some extent. The attention weights indicate the importance of a cough to an assessment task. Figure 5 uses the healthy *v.s.* unhealthy task as an example and visualizes the attention weights distribution of all subjects in the second dataset. Overall, the majority of the weights are close to zero and only a small fraction of the cough snippets have weights that are larger than 0.2 (5.0±2.4 per person on average, 4.0%±2.0%). This indicates that only a few coughs were picked by the MIL framework and effective for the assessment task.

As an anecdotal example, we randomly picked two subjects that were correctly classified in the healthy *v.s.* unhealthy task, one belonging to the healthy group (noted as P1, with FEV1 $\leq$ 50% predicted FEV1) and the other belonging to the unhealthy group (noted as P2, with FEV1 $\geq$ 50% predicted FEV1). P1 had 106 cough sound snippets and P2 had 155. We inspected the parameters for all snippets of the two subjects. The left side of Figure 6 shows the weights of these snippets. To further understand what coughs were selected for each subject, we

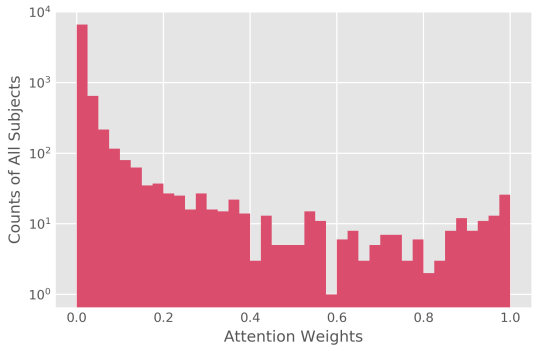
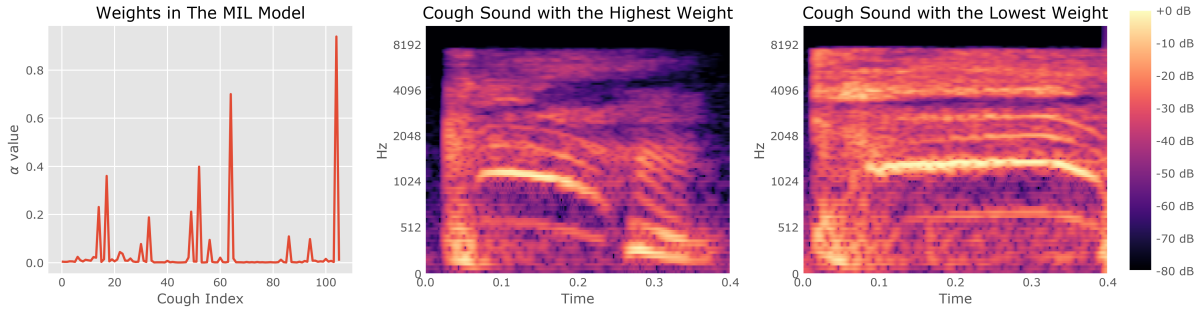
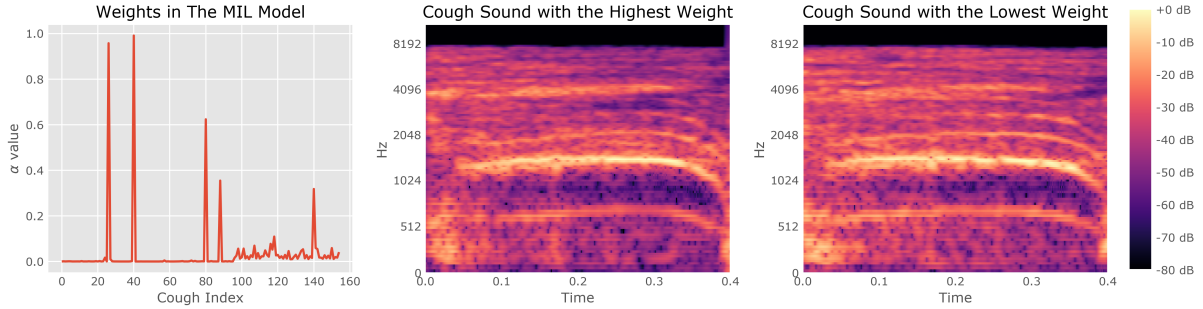


Fig. 5. The histogram of attention weights of all subjects. Note that the y-axis is in log scale.



(a) A sample subject with FEV1 ≤ 50% predicted FEV1 (unhealthy)



(b) A sample subject with FEV1 ≥ 50% predicted FEV1 (healthy)

Fig. 6. Both figures have the same layout. Left) The attention weight α of each cough. Middle) The mel-spectrogram of the cough sound with the highest weight (#104 for a and #40 for b). Right) The mel-spectrogram of the cough sound with the lowest weight (#44 for a and #41 for b).

picked one cough sound among those with the highest weights, and one cough sound among those with the lowest weights. Figure 6a shows the mel-spectrograms of the two coughs for P1, and Figure 6b shows the ones for P2. The comparison between the same subject's two coughs shows interesting findings. For P1 (unhealthy), the two cough sounds were quite different. The cough with the highest weight had an obvious short interruption at 0.25 s, while the cough with the lowest weight did not show such an interruption. In contrast, for P2 (healthy), the two coughs had similar spectrograms. This indicates that for the subject who had severe lung health issues, the model might be able to select the coughs that were potentially more related to the symptoms. However, we did not see a similar selection for the subject who did not have severe lung health issues. This is supported by recent medical studies' findings that voluntary coughs have different characteristics based on how much they are correlated with the lung condition [15]. We conducted an informal unstructured interview with a medical expert to ask about such an observation. The expert confirmed that if a subject had lower-than-normal FEV1, which suggests potential breathing obstruction, a cough could potentially be affected by the narrow or inflamed airways. These results validate the ability of our method to provide high-level interpretability.

6 DISCUSSION

In this section, we discuss the problem of the public cough-related datasets (Section 6.1), the relationship between the cough detection task and the assessment tasks (Section 6.3), as well as the future of leveraging our method for diagnosis in real settings (Section 6.3). We also reflected on important limitations in this work (Section 6.4).

6.1 Upcoming Cough Datasets

Existing public large-scale cough sound datasets are still limited nowadays. In our work, since the available cough data in ESC-50 and Freesound is quite small, the only dataset that contains rich cough audio is the Google AudioSet (also with the manual labels from FluSense [5]). After the pre-processing described in Section 4.3, we had around 15 thousand cough audio from the three public datasets. This is not a big number for training very deep network architecture. Although we have leveraged a number of augmentation techniques to mitigate this issue, it can be expected that there is still great room for cough detection improvement when more cough data becomes available. Due to the COVID-19, there have been emerging studies and mobile applications that collected cough sounds for various usage. For example, University of Cambridge recently released a mobile app that collected COVID-19 coughs and other sounds [2]. Therefore, it is promising that in the near future, there will be more and more large-scale cough detection datasets becoming publicly available. Our method can greatly benefit from the further richness of the data.

6.2 The Relationship between Cough Detection and Lung Health Assessment

Our findings in Section 5.2 indicate that transferring the cough detection model to the lung health assessment models did significantly improve the performance. This reflects the interesting relationship between these two types of tasks. For cough detection, the goal of the model is to extract an effective representation that can differentiate cough against other noise. While for the assessment task, the goal is to extract a representation from cough audio that can help distinguish different lung health conditions (*e.g.*, obstructive *v.s.* non-obstructive). Although targeted at different tasks, these goals have overlap on processing and extracting feature vectors from cough sounds. The features extracted for cough detection are also useful to distinguish different lung health conditions. This might be able to explain why initializing the assessment models with the cough detection model can produce much better results.

6.3 Computer-Aided Diagnosis

Our current method is able to provide limited interpretability: the model can only identify the important coughs for these tasks, which might be able to help reduce expert's effort of going through audio materials and pin-point those "suspicious" sounds. This is consistent with medical studies [15]. According to our conversation with the medical expert (mentioned in Section 5.2.3), the identified cough in Figure 6a could be helpful for diagnosis. However, for the subject without severe lung health issues, it is hard for an expert to confidently claim the outcome by just seeing that the coughs with different weights are similar. *"Simply showing these figures* [(Figure 6b)] *is not convincing enough.* They would worry about the false-negative cases. *"Even if you tell me that all coughs are similar and the model says negative, I would still need more evidence to offer a valid outcome. Otherwise, it may be easy to miss a patient who actually has some pulmonary diseases."* It still remains as an open question of how such a method can further assist experts for diagnosis. In addition to further improving the accuracy of a machine learning model, another potential solution is to further investigate the relationship of cough sounds with more detailed physiological symptoms such as the amount of phlegm or sputum, or lung function parameters such as FEV1, FVC, and FEV1/FVC ratio [76].

6.4 Limitation and Future Work

There are a few limitations in our current work. First, one of the advantages of an end-to-end audio model is to simplify the software and hardware computation so that an on-device pipeline can become feasible. However, in this paper, our focus is on achieving better generalizability of the cough detection model, and better accuracy of the lung health assessment models. Although we used MobileNet to reduce the parameter number, we did not try more advanced models (*e.g.*, the latest MobileNet-v3 [38] and FBNet-v2 [83]), or conduct further investigation

techniques such as quantization [36]. In the future, we plan to explore more model compression models and evaluate the model's performance on mobile and wearable devices. Second, there are some generalizability aspects of the cough detection model that we did not explore. For example, how does the model perform when tested on the audio collected from smartwatches? Moreover, the generalizability of the assessment models was not evaluated on another new clinical dataset. These limitations are mainly due to the restriction of the data. To address this issue, we plan to collect more cough data using devices in multiple form factors. We also plan to conduct more studies in collaboration with our partner pulmonologists. Third, the interpretability of our current method and how it can help experts for diagnosis needs deeper investigation. As mentioned in Section 6.3, understanding experts' needs can provide valuable guidance for the design of future models. Currently, we only had a brief informal interview with one medical expert. We plan to conduct a formal structured or semi-structured interview with more expects in the future.

7 CONCLUSION

In this work, we propose Listen2Cough for a series of cough-related tasks. We first develop a new end-to-end deep learning architecture based on MobileNet and a number of augmentation techniques to improve the generalizability of a cough detection model. Moreover, to address the issue of the limited dataset in prior work on lung health assessment, we further propose to connect the cough detection task with the health assessment tasks by transfer learning in a MIL framework with the attention mechanism. Our experiments show that our cough detection model significantly outperforms the baseline by 5.8% on F1 score when applied on a clean dataset and 8.4% when applied on a noisy in-the-wild dataset. In addition to the good results on cough detection, our transferred lung assessment model also outperforms the baseline by 10.7%, 6.3%, and 3.7% for healthy *v.s.* Unhealthy, obstructive *v.s.* non-obstructive, and COPD *v.s.* asthma, respectively. Moreover, our investigation shows that the attention relative weights in the MIL framework can identify important coughs for the assessment tasks, which provides interpretability that can potentially help experts for better diagnosis in the future.

ACKNOWLEDGMENTS

We want to acknowledge Hujun Cui for the phone and watch application development, Leonardo Jimenez Rodriguez, Yoshiya Hirase and Sharath Chandrashekhara for preparing the back-end servers, FigureEight for data annotation, Dr. Erin Blackstock for data collection supervision, Dr. Christopher Fanta for clinical feedback, and the participants for volunteering in the study.

REFERENCES

[1] 2014. The Cost of Lung Disease. Lung Health Institute. https://lunginstitute.com/blog/the-cost-of-lung-disease/
[2] 2020. Covid-19 Sounds App. https://www.covid-19-sounds.org/en/
[3] 2020. Global Initiative for Chronic Obstructive Lung Disease - Global Initiative for Chronic Obstructive Lung Disease. https://goldcopd.org/
[4] 2020. Lung Health & Diseases. American Lung Association. https://www.lung.org/lung-health-diseases
[5] Forsad Al Hossain, Andrew A Lover, George A Corey, Nicholas G Reich, and Tauhidur Rahman. 2020. FluSense: a contactless syndromic surveillance platform for influenza-like illness in hospital waiting areas. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–28.
[6] Justice Amoh and Kofi Odame. 2015. DeepCough: A deep convolutional neural network in a wearable cough detection system. In *2015 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 1–4.
[7] Justice Amoh and Kofi Odame. 2016. Deep neural networks for identifying cough sounds. *IEEE transactions on biomedical circuits and systems* 10, 5 (2016), 1003–1011.
[8] Yusuf A Amrulloh, Udantha R Abeyratne, Vinayak Swarnkar, Rina Triasih, and Amalia Setyati. 2015. Automatic cough segmentation from non-contact sound recordings in pediatric wards. *Biomedical Signal Processing and Control* 21 (2015), 126–136.
[9] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. Soundnet: Learning sound representations from unlabeled video. In *Advances in neural information processing systems*. 892–900.

[10] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[11] Filipe Barata, Kevin Kipfer, Maurice Weber, Peter Tinschert, Elgar Fleisch, and Tobias Kowatsch. 2019. Towards device-agnostic mobile cough detection with convolutional neural networks. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE, 1–11.

[12] Charles Bergeron, Gregory Moore, Jed Zaretzki, Curt M Breneman, and Kristin P Bennett. 2011. Fast bundle algorithm for multiple-instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 6 (2011), 1068–1079.

[13] Hylan A Bickerman and Sylvia E Itkin. 1958. The effect of a new bronchodilator aerosol on the air flow dynamics of the maximal voluntary cough of patients with bronchial asthma and pulmonary emphysema. *Journal of chronic diseases* 8, 5 (1958), 629–636.

[14] SS Birring, T Fleming, S Matos, AA Raj, DH Evans, and ID Pavord. 2008. The Leicester Cough Monitor: preliminary validation of an automated cough detection system in chronic cough. *European Respiratory Journal* 31, 5 (2008), 1013–1018.

[15] James C Borders, Alexandra E Brandimore, and Michelle S Troche. 2020. Variability of Voluntary Cough Airflow in Healthy Adults and Parkinson's Disease. *Dysphagia* (2020), 1–7.

[16] Marc-André Carbonneau, Veronika Cheplygina, Eric Granger, and Ghyslain Gagnon. 2018. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognition* 77 (2018), 329–353.

[17] Marc-André Carbonneau, Eric Granger, Yazid Attabi, and Ghyslain Gagnon. 2017. Feature learning from spectrograms for assessment of personality traits. *IEEE Transactions on Affective Computing* (2017).

[18] Daniel B Chamberlain, Rahul Kodgule, and Richard Ribón Fletcher. 2016. A mobile platform for automated screening of asthma and chronic obstructive pulmonary disease. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 5192–5195.

[19] Soujanya Chatterjee, Md Mahbubur Rahman, Tousif Ahmed, Nazir Saleheen, Ebrahim Nemati, Viswam Nathan, Korosh Vatanparvar, and Jilong Kuang. 2020. Assessing Severity of Pulmonary Obstruction from Respiration Phase-Based Wheeze-Sensing Using Mobile Sensors. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376444

[20] Sneha Chaudhari, Gungor Polatkan, Rohan Ramanath, and Varun Mithal. 2019. An attentive survey of attention models. *arXiv preprint arXiv:1904.02874* (2019).

[21] Qian Cheng, Joshua Juen, Shashi Bellam, Nicholas Fulara, Deanna Close, Jonathan C Silverstein, and Bruce Schatz. 2017. Predicting pulmonary function from phone sensors. *Telemedicine and e-Health* 23, 11 (2017), 913–919.

[22] Veronika Cheplygina, Lauge Sørensen, David MJ Tax, Jesper Holst Pedersen, Marco Loog, and Marleen de Bruijne. 2014. Classification of COPD with multiple instance learning. In *2014 22nd International Conference on pattern recognition*. IEEE, 1508–1513.

[23] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*. 3504–3512.

[24] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. 2016. Weakly supervised object localization with multi-fold multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence* 39, 1 (2016), 189–203.

[25] Wei Dai, Chia Dai, Shuhui Qu, Juncheng Li, and Samarjit Das. 2017. Very deep convolutional neural networks for raw waveforms. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 421–425.

[26] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[27] Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence* 89, 1-2 (1997), 31–71.

[28] Mohammad Ebrahimpour, Timothy Shea, Andreea Danielescu, David Noelle, and Chris Kello. 2020. End-to-End Auditory Object Recognition via Inception Nucleus. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 146–150.

[29] Eduardo Fonseca, Manoj Plakal, Frederic Font, Daniel PW Ellis, Xavier Favory, Jordi Pons, and Xavier Serra. 2018. General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline. *arXiv preprint arXiv:1807.09902* (2018).

[30] Wei Gao, Wuping Bao, and Xin Zhou. 2019. Analysis of cough detection index based on decision tree and support vector machine. *Journal of Combinatorial Optimization* 37, 1 (2019), 375–384.

[31] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 776–780.

[32] Aurélien Géron. 2019. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

[33] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep Sparse Rectifier Neural Networks. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*. 315–323.

[34] Mayank Goel, Elliot Saba, Maia Stiber, Eric Whitmire, Josh Fromm, Eric C. Larson, Gaetano Borriello, and Shwetak N. Patel. 2016. Spirocall: Measuring lung function over a phone call. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*.

ACM, New York, NY, USA, 5675–5685. https://doi.org/10.1145/2858036.2858401
[35] Siddharth Gupta, Peter Chang, Nonso Anyigbo, and Ashutosh Sabharwal. 2011. mobileSpiro: accurate mobile spirometry for self-management of asthma. In *Proceedings of the First ACM Workshop on Mobile Systems, Applications, and Services for Healthcare*. 1–6.
[36] Song Han, Huizi Mao, and William J Dally. 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149* (2015).
[37] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
[38] Andrew Howard, Mark Sandler, Bo Chen, Weijun Wang, Liang Chieh Chen, Mingxing Tan, Grace Chu, Vijay Vasudevan, Yukun Zhu, and Ruoming and Pang. 2019. Searching for MobileNetV3. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*.
[39] Jonathan J Huang and Juan Jose Alvarado Leanos. 2018. Aclnet: efficient end-to-end audio classification cnn. *arXiv preprint arXiv:1811.06669* (2018).
[40] MA Huckvale and András Beke. 2017. It sounds like you have a cold! Testing voice features for the Interspeech 2017 Computational Paralinguistics Cold Challenge. International Speech Communication Association (ISCA).
[41] Maximilian Ilse, Jakub M Tomczak, and Max Welling. 2018. Attention-based deep multiple instance learning. *arXiv preprint arXiv:1802.04712* (2018).
[42] Ali Imran, Iryna Posokhova, Haneya N Qureshi, Usama Masood, Sajid Riaz, Kamran Ali, Charles N John, and Muhammad Nabeel. 2020. AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. *arXiv preprint arXiv:2004.01275* (2020).
[43] Marco Jeub, Magnus Schafer, and Peter Vary. 2009. A binaural room impulse response database for the evaluation of dereverberation algorithms. In *2009 16th International Conference on Digital Signal Processing*. IEEE, 1–5.
[44] Joshua Juen, Qian Cheng, Valentin Prieto-Centurion, Jerry A Krishnan, and Bruce Schatz. 2014. Health monitors for chronic disease by gait analysis with mobile phones. *Telemedicine and e-Health* 20, 11 (2014), 1035–1041.
[45] Joshua Juen, Qian Cheng, and Bruce Schatz. 2015. A natural walking monitor for pulmonary patients using mobile phones. *IEEE Journal of biomedical and health informatics* 19, 4 (2015), 1399–1405.
[46] Simon Kornblith, Jonathon Shlens, and Quoc V Le. 2019. Do better imagenet models transfer better?. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2661–2671.
[47] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90.
[48] Jordi Laguarta, Ferran Hueto, and Brian Subirana. 2020. COVID-19 Artificial Intelligence Diagnosis using only Cough Recordings. *IEEE Open Journal of Engineering in Medicine and Biology* (2020).
[49] Gierad Laput, Karan Ahuja, Mayank Goel, and Chris Harrison. 2018. Ubicoustics: Plug-and-play acoustic activity recognition. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. 213–224.
[50] Eric C. Larson, Mayank Goel, Gaetano Boriello, Sonya Heltshe, Margaret Rosenfeld, and Shwetak N. Patel. 2012. SpiroSmart: using a microphone to measure lung function on a mobile phone. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing - UbiComp '12*. ACM Press, New York, New York, USA, 280. https://doi.org/10.1145/2370216.2370261
[51] Shasha Le and Weiping Hu. 2013. Cough sound recognition based on Hilbert marginal spectrum. In *2013 6th International Congress on Image and Signal Processing (CISP)*, Vol. 3. IEEE, 1346–1350.
[52] John Boaz Lee, Ryan A Rossi, Sungchul Kim, Nesreen K Ahmed, and Eunyee Koh. 2019. Attention models in graphs: A survey. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 13, 6 (2019), 1–25.
[53] Jia-Ming Liu, Mingyu You, Guo-Zheng Li, Zheng Wang, Xianghuai Xu, Zhongmin Qiu, Wenjia Xie, Chao An, and Sili Chen. 2013. Cough signal recognition with gammatone cepstral coefficients. In *2013 IEEE China Summit and International Conference on Signal and Information Processing*. IEEE, 160–164.
[54] Jia-Ming Liu, Mingyu You, Zheng Wang, Guo-Zheng Li, Xianghuai Xu, and Zhongmin Qiu. 2014. Cough detection using deep neural networks. In *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 560–563.
[55] Gianluca Maguolo, Michelangelo Paci, Loris Nanni, and Ludovico Bonan. 2019. Audiogmenter: a MATLAB Toolbox for Audio Data Augmentation. *arXiv preprint arXiv:1912.05472* (2019).
[56] David M Mannino, Earl S Ford, and Stephen C Redd. 2003. Obstructive and restrictive lung disease and markers of inflammation: data from the Third National Health and Nutrition Examination. *The American journal of medicine* 114, 9 (2003), 758–762.
[57] Sergio Matos, Surinder S Birring, Ian D Pavord, and H Evans. 2006. Detection of cough signals in continuous audio recordings using hidden Markov models. *IEEE Transactions on Biomedical Engineering* 53, 6 (2006), 1078–1083.
[58] Puja Mehta, Daniel F McAuley, Michael Brown, Emilie Sanchez, Rachel S Tattersall, Jessica J Manson, HLH Across Speciality Collaboration, *et al.* 2020. COVID-19: consider cytokine storm syndromes and immunosuppression. *Lancet (London, England)* 395, 10229 (2020), 1033.
[59] Jesús Monge-Álvarez, Carlos Hoyos-Barceló, Paul Lesso, and Pablo Casaseca-de-la Higuera. 2018. Robust detection of audio-cough events using local Hu moments. *IEEE journal of biomedical and health informatics* 23, 1 (2018), 184–196.
[60] Viswam Nathan, Korosh Vatanparvar, Md Mahbubur Rahman, Ebrahim Nemati, and Jilong Kuang. 2019. Assessment of chronic pulmonary disease patients using biomarkers from natural speech recorded by mobile devices. In *2019 IEEE 16th International Conference*

on Wearable and Implantable Body Sensor Networks (BSN). IEEE, 1–4.

[61] Ebrahim Nemati, Md Juber Rahman, Korosh Vatanparvar, Viswam Nathan, and Jilong Kuang. 2020. Estimation of the Lung Function Using Acoustic Features of the Voluntary Cough. *42nd Annual International Conferences of the IEEE Engineering in Medicine and Biology Society, EMBC 2020*.

[62] Ebrahim Nemati, Md Mahbubur Rahman, Viswam Nathan, and Jilong Kuang. 2018. Private audio-based cough sensing for in-home pulmonary assessment using mobile devices. In *EAI International Conference on Body Area Networks*. Springer, 221–232.

[63] Tuomas Oikarinen, Karthik Srinivasan, Olivia Meisner, Julia B Hyman, Shivangi Parmar, Adrian Fanucci-Kiss, Robert Desimone, Rogier Landman, and Guoping Feng. 2019. Deep convolutional network for animal sound classification and source attribution using dual audio recordings. *The Journal of the Acoustical Society of America* 145, 2 (2019), 654–662.

[64] Nikolaos Pappas and Andrei Popescu-Belis. 2014. Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis. In *Proceedings of the 2014 Conference on Empirical Methods In Natural Language Processing (EMNLP)*. 455–466.

[65] Karol J Piczak. 2015. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*. 1015–1018.

[66] Liam Polley, Nurman Yaman, Liam Heaney, Chris Cardwell, Eimear Murtagh, John Ramsey, Joseph MacMahon, Richard W Costello, and Lorcan McGarvey. 2008. Impact of cough across different chronic respiratory diseases: comparison of two cough-specific health-related quality of life questionnaires. *Chest* 134, 2 (2008), 295–302.

[67] Renard Xaviero Adhi Pramono, Syed Anas Imtiaz, and Esther Rodriguez-Villegas. 2016. A cough-based algorithm for automatic diagnosis of pertussis. *PloS one* 11, 9 (2016), e0162128.

[68] Speech Processing. 2008. Transmission and Quality Aspects (STQ); Speech Quality Performance in the Presence of Background Noise; Part 1: Background Noise Simulation Technique and Background Noise Database. *ETSI EG* 202 (2008), 396–1.

[69] Gwénolé Quellec, Mathieu Lamard, Michael D Abràmoff, Etienne Decencière, Bruno Lay, Ali Erginay, Béatrice Cochener, and Guy Cazuguel. 2012. A multiple-instance learning framework for diabetic retinopathy screening. *Medical image analysis* 16, 6 (2012), 1228–1240.

[70] Colin Raffel and Daniel PW Ellis. 2015. Feed-forward networks with attention can solve some long-term memory problems. *arXiv preprint arXiv:1512.08756* (2015).

[71] Md Juber Rahman, Ebrahim Nemati, Md Mahbubur Rahman, Viswam Nathan, Korosh Vatanparvar, and Jilong Kuang. 2020. Automated assessment of pulmonary patients using heart rate variability from everyday wearables. *Smart Health* 15 (2020), 100081.

[72] Md Mahbubur Rahman, Tousif Ahmed, Ebrahim Nemati, Viswam Nathan, Korosh Vatanparvar, Erin Blackstock, and Jilong Kuang. 2020. ExhaleSense: Detecting High Fidelity Forced Exhalations to Estimate Lung Obstruction on Smartphones. In *2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 1–10. https://doi.org/10.1109/PerCom45495.2020.9127355

[73] Jose F Ruiz-Muñoz, Mauricio Orozco Alzate, and Germán Castellanos-Domínguez. 2015. Multiple instance learning-based birdsong classification using unsupervised recording segmentation. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

[74] Keum San Chun, Viswam Nathan, Korosh Vatanparvar, Ebrahim Nemati, Md Mahbubur Rahman, Erin Blackstock, and Jilong Kuang. 2020. Towards Passive Assessment of Pulmonary Function from Natural Speech Recorded Using a Mobile Phone. In *2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 1–10.

[75] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4510–4520.

[76] Roneel V Sharan, Udantha R Abeyratne, Vinayak R Swarnkar, Scott Claxton, Craig Hukins, and Paul Porter. 2018. Predicting spirometry readings using cough sound features and regression. *Physiological measurement* 39, 9 (2018), 095001.

[77] Neeraj Sharma, Prashant Krishnan, Rohit Kumar, Shreyas Ramoji, Srikanth Raj Chetupalli, Prasanta Kumar Ghosh, Sriram Ganapathy, *et al*. 2020. Coswara–A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis. *arXiv preprint arXiv:2005.10548* (2020).

[78] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[79] Jaclyn Smith and Ashley Woodcock. 2006. Cough and its importance in COPD. *International journal of chronic obstructive pulmonary disease* 1, 3 (2006), 305.

[80] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. 2016. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261* (2016).

[81] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. 2013. The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings. In *Proceedings of Meetings on Acoustics ICA2013*, Vol. 19. Acoustical Society of America, 035081.

[82] Tong Tong, Robin Wolz, Qinquan Gao, Ricardo Guerrero, Joseph V Hajnal, Daniel Rueckert, Alzheimer's Disease Neuroimaging Initiative, *et al*. 2014. Multiple instance learning for classification of dementia in brain MRI. *Medical image analysis* 18, 5 (2014), 808–818.

[83] Alvin Wan, Xiaoliang Dai, Peizhao Zhang, Zijian He, Yuandong Tian, Saining Xie, Bichen Wu, Matthew Yu, Tao Xu, Kan Chen, Peter Vajda, and Joseph E. Gonzalez. 2020. FBNetV2: Differentiable Neural Architecture Search for Spatial and Channel Dimensions. In *arXiv*.

arXiv:2004.05565

[84] Yan Xu, Jun-Yan Zhu, I Eric, Chao Chang, Maode Lai, and Zhuowen Tu. 2014. Weakly supervised histopathology cancer image segmentation and classification. *Medical image analysis* 18, 3 (2014), 591–604.

[85] Aina M Yañez, Dolores Guerrero, Rigoberto Pérez de Alejo, Francisco Garcia-Rio, Jose Luis Alvarez-Sala, Miriam Calle-Rubio, Rosa Malo de Molina, Manuel Valle Falcones, Piedad Ussetti, Jaume Sauleda, *et al.* 2012. Monitoring breathing rate at home allows early identification of COPD exacerbations. *Chest* 142, 6 (2012), 1524–1529.

[86] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* (2017).

[87] Zhi-Hua Zhou. 2004. Multi-instance learning: A survey. *Department of Computer Science & Technology, Nanjing University, Tech. Rep* 1 (2004).

[88] Zhi-Hua Zhou, Kai Jiang, and Ming Li. 2005. Multi-instance learning based web mining. *Applied intelligence* 22, 2 (2005), 135–147.

[89] Chunmei Zhu, Lianfang Tian, Xiangyang Li, Hongqiang Mo, and Zeguang Zheng. 2013. Recognition of cough using features improved by sub-band energy transformation. In *2013 6th International Conference on Biomedical Engineering and Informatics*. IEEE, 251–255.

[90] Fatma Zubaydi, Assim Sagahyroon, Fadi Aloul, and Hasan Mir. 2017. MobSpiro: Mobile based spirometry for detecting COPD. In *2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC)*. IEEE, 1–4.