Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data

Chloë Brown∗
clb76@cam.ac.uk
University of Cambridge

Jagmohan Chauhan∗
jc2161@cam.ac.uk
University of Cambridge

Andreas Grammenos†
ag926@cl.cam.ac.uk
University of Cambridge

Jing Han∗
jh2298@cam.ac.uk
University of Cambridge

Apinan Hasthanasombat∗
ah953@cam.ac.uk
University of Cambridge

Dimitris Spathis∗
ds806@cam.ac.uk
University of Cambridge

Tong Xia∗
xiatongxt028@gmail.com
University of Cambridge

Pietro Cicuta
pc245@cam.ac.uk
University of Cambridge

Cecilia Mascolo
cm542@cam.ac.uk
University of Cambridge

1 INTRODUCTION

Audio signals generated by the human body (e.g., sighs, breathing, heart, digestion, vibration sounds) have routinely been used by clinicians as diagnostic or progression indicators for diseases and disease onset. However, until recently, such signals were usually collected through manual auscultation at scheduled visits. Research has now started to use digital technology to gather bodily sounds (e.g., from digital stethoscopes) for cardiovascular or respiratory examination, which could then be used for automatic analysis. Some initial work shows promise in detecting diagnostic signals of COVID-19 from voice and coughs. In this paper we describe our data analysis over a large-scale crowdsourced dataset of respiratory sounds collected to aid diagnosis of COVID-19. We use coughs and breathing to understand how discernible COVID-19 sounds are from those in asthma or healthy controls. Our results show that even a simple binary machine learning classifier is able to classify correctly healthy and COVID-19 sounds. We also show how we distinguish a user who tested positive for COVID-19 and has a cough from a healthy user with cough, and users who tested positive for COVID-19 and have a cough from users with asthma and a cough. Our models achieve an AUC above 70% across all tasks. Clearly these results are preliminary and only scratch the surface of the possibilities of the exploitation of this type of data and audio-based machine learning. This work opens the door to further investigation of how automatically analysed respiratory patterns could be used as pre-screening signals to aid COVID-19 diagnosis.

KEYWORDS
COVID-19, Crowdsourcing Platform, Audio Analysis, Coughing, Breathing

∗Ordered alphabetically, equal contribution.
†Also with Alan Turing Institute.

ACM Reference Format:
Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Jing Han, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, and Cecilia Mascolo. 2020. Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data. In Proceedings of (pre-print). ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

ABSTRACT

Audio signals generated by the human body (e.g., sighs, breathing, heart, digestion, vibration sounds) have routinely been used by clinicians as diagnostic or progression indicators for diseases and disease onset. However, until recently, such signals were usually collected through manual auscultation at scheduled visits. Research has now started to use digital technology to gather bodily sounds (e.g., from digital stethoscopes) for cardiovascular or respiratory examination, which could then be used for automatic analysis. Some initial work shows promise in detecting diagnostic signals of COVID-19 from voice and coughs. In this paper we describe our data analysis over a large-scale crowdsourced dataset of respiratory sounds collected to aid diagnosis of COVID-19. We use coughs and breathing to understand how discernible COVID-19 sounds are from those in asthma or healthy controls. Our results show that even a simple binary machine learning classifier is able to classify correctly healthy and COVID-19 sounds. We also show how we distinguish a user who tested positive for COVID-19 and has a cough from a healthy user with cough, and users who tested positive for COVID-19 and have a cough from users with asthma and a cough. Our models achieve an AUC above 70% across all tasks. Clearly these results are preliminary and only scratch the surface of the possibilities of the exploitation of this type of data and audio-based machine learning. This work opens the door to further investigation of how automatically analysed respiratory patterns could be used as pre-screening signals to aid COVID-19 diagnosis.

KEYWORDS
COVID-19, Crowdsourcing Platform, Audio Analysis, Coughing, Breathing

∗Ordered alphabetically, equal contribution.
†Also with Alan Turing Institute.

ACM Reference Format:
Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Jing Han, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, and Cecilia Mascolo. 2020. Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data. In Proceedings of (pre-print). ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Audio signals generated by the human body (e.g., sighs, breathing, heart, digestion, vibration sounds) have often been used by clinicians and clinical researchers as diagnostic or progression indicators for diseases and disease onset. However, until recently, such signals were usually collected through manual auscultation at scheduled visits. Research has now started to use digital technology to gather bodily sounds (e.g., digital stethoscopes) and run automatic analysis on the data [24], for example for wheeze detection in asthma [18, 23]. Researchers have also been piloting the use of human voice to assist early diagnosis of a variety of illnesses: Parkinson’s disease correlates with softness of speech (resulting from lack of coordination of the vocal muscles) [6, 12], voice frequency with coronary artery disease (hardening of the arteries which may affect voice production) [19] or vocal tone, pitch, rhythm, rate, and volume correlate with invisible injuries such as post-traumatic stress disorder [5], traumatic brain injury and psychiatric conditions [13]. The use of human-generated audio as a biomarker for various illnesses offers enormous potential for early diagnosis, as well as for affordable solutions which could be rolled out to the masses if embedded in commodity devices. This is even more true if such solutions could monitor individuals throughout their daily lives in an unobtrusive way.

Recent work has started exploring how respiratory sounds (e.g., coughs, breathing and voice) collected by devices from COVID-19 positively tested patients in hospital differ from sounds from healthy people. In [16] digital stethoscope data from lung auscultation is used as a diagnostic signal for COVID-19; in [17] a study of detection of coughs related to COVID-19 collected with phones is presented using a cohort of 48 COVID-19 tested patients versus other pathological coughs on which an AI engine is trained. In [14] speech recordings from COVID-19 hospital patients are analyzed...
to categorize automatically the health state of patients. Our work contains an exploration of using human respiratory sounds as diagnostic markers for COVID-19 in crowdsourced, uncontrolled data. Specifically, this paper describes our preliminary findings over a subset of our dataset currently being crowdsourced worldwide. The dataset was collected through an app (Android and Web) that asked volunteers for samples of their voice, coughs and breathing as well as their medical history and symptoms. The app also asks if the user has tested positive for COVID-19. To date, we have collected on the order of 10,000 samples from about 7000 unique users. While other efforts exist that collect similar data, they are often either limited in scope (e.g., collect only coughs [1, 2]) or in scale (e.g., collect smaller samples in a specific region or hospital). This is, to our knowledge, the largest uncontrolled, crowdsourced data collection of COVID-19 related sounds worldwide. In addition, the mobile app gathers data from single individuals up to every two days, allowing for potential tracking of disease progression. This is also a unique feature of our collected dataset. Section 3 contains a more detailed description of the data. In the paper we analyze a subset of our data as described in Section 3.3 and show some preliminary evidence that cough and breathing sounds could contain diagnostic signals to discriminate COVID-19 users from healthy ones: we further compare COVID-19 positive user coughs with healthy coughs as well as those from users with asthma.

More precisely the contributions of this paper are:

- We describe our COVID-19 sound collection framework through apps, and the types of sounds harvested through crowdsourcing.
- We illustrate the large-scale dataset being gathered. To date, this is the largest being collected and among the most inclusive in terms of types of sounds. It contains sounds from about 7000 unique users (more than 200 of whom reported they have tested positive for COVID-19 recently).
- We describe our initial findings around the discriminatory power of coughs and breathing sounds for COVID-19. We construct three binary tasks, one aimed at distinguishing COVID-19 positive users from healthy users; one aimed at distinguishing COVID-19 positive users who have a cough from healthy users who have a cough; and one aimed at distinguishing COVID-19 positive users with a cough from users with asthma who declared a cough. The precise results show that our performance remains above 70% AUC for all tasks. In particular, we show that even a simple binary machine learning classifier is able to classify correctly healthy and COVID-19 sounds with an Area Under Curve (AUC) of 72%. When trying to distinguish a user who tested positive for COVID-19 and has a cough from a healthy user with a cough, our classifier reaches an AUC of 75%, while if we try to distinguish users who tested positive for COVID-19 and have a cough from users with asthma and a cough we achieve an AUC of 76%.
- We test how audio data augmentation can be used to improve the recall performance of some of our tasks with less data.
- We thoroughly discuss our results and their potential, and illustrate a number of future directions for our analysis and for sound-based diagnostics in the context of COVID-19, which could open the door to COVID-19 pre-screening and progression detection.

2 Motivation and Related Work

Researchers have long recognised the utility of sound as a possible indicator of behaviour and health. Purpose-built external microphones have been used to detect sound from the heart or the lungs using stethoscopes, for example. However, these often require listening and interpretation by highly skilled clinicians, and are recently and rapidly being substituted by different technologies such as a variety of imaging techniques (e.g., MRI, sonography), for which analysis and interpretation is easier. However, recent trends in automated audio interpretation and modeling has the potential to reverse this trend and offer sound as the cheap and easily distributable alternative.

More recently the microphone on commodity devices such as smartphones and wearables have been exploited for sound analysis. In [8] the audio from the microphone is used to understand the user context and this information is aggregated to make up a view of the ambience of places around a city. In Emotionsense [26], the phone microphone is used as a sensor for users’ emotion detection in-the-wild through Gaussian mixture models. In [22] authors analyze sounds emitted while the user is sleeping to signal sleep apnea episodes. Similar works have also used sound to detect asthma and wheezing [18, 23].

Machine learning methods have been devised to recognize and diagnose respiratory diseases from sounds [24] and more specifically coughs: [4] uses convolutional neural networks (CNNs) to detect cough within ambient audio, and diagnose three potential illnesses (bronchitis, bronchiolitis and pertussis) based on their unique audio characteristics.

Clinical work has concentrated on using voice analysis for specific diseases: for example, in Parkinson’s disease, microphone and laryngograph equipment have been used to detect the softness of speech resulting from lack of coordination over the vocal muscles [6, 12]. Voice features have also been used to diagnose bipolar disorder [13]; and to correlate tone, pitch, rhythm, rate, and volume with signs of invisible injuries like post traumatic stress disorder [5], traumatic brain injury and depression. Voice frequency has been linked to coronary artery disease (resulting from the hardening of the arteries which may affect voice production) [19]. Companies such as Israeli-based Beyond Verbal and the Mayo Clinic have been indicated in the press releases as piloting these approaches.

Recently, with the advent of COVID-19, researchers have started to explore if respiratory sounds could be diagnostic [10]. In [16] digital stethoscope data from lung auscultation is used as a diagnostic signal for COVID-19. In [17] a study of detection of coughs related to COVID-19 is presented using a cohort of 48 COVID-19 tested patients versus other pathology coughs on which an AI model is trained. In [14] speech recordings from COVID-19 patients are analyzed to automatically categorize the health state of patients from four aspects, including the severity of illness, sleep quality, fatigue, and anxiety.

Our work differs from these works, as we use an entirely crowdsourced dataset, for which we must trust that the ground truth is what the users state (in terms of symptoms and COVID-19 testing
status); we further have to overcome the challenges of data coming from different phones and microphones as well as in possibly very different environments. Other crowdsourced approaches of this kind are starting to emerge: in [28] a web form to gather sound data is presented which collected about 570 samples but does not report any COVID-19 detection analysis. Our app collected samples from more than 7000 unique users with more than 200 positively tested for COVID-19 and allows for users to go back to the app after a few days to report progression and give another sample. We report our preliminary findings which suggest that sounds could be used in some forms to inform automatic COVID-19 screening.

3 DATA COLLECTION

This section describes the data collection framework and some properties of the gathered data. We further describe in detail the subset of the data which we use for the analysis in this paper. We note that the data collection and study have been approved by the Ethics Committee of the Department of Computer Science and Technology at the University of Cambridge.

3.1 Our data collection apps

Our crowdsourced data gathering framework is composed of a web-based app and an Android app. Our iOS app has just been released by Apple but we have not yet collected data from it. Most of the features of the web and mobile apps are similar. A user is asked to input their age and gender as well as their medical history and whether they are in hospital. The users then input their symptoms (if any) and their respiratory sounds: they are asked to cough three times, to breathe deeply through their mouth three to five times and to read a short sentence appearing on the screen three times. Finally, the users are asked if they have been tested for COVID-19 and a location sample is also gathered. Figure 1 illustrates the screens of the Android app collecting coughs and symptoms. In addition, the Android (and iOS) app prompts the user to input further sounds and symptoms every two days, providing a unique opportunity to study the progression of user health based on sounds. The data flows encrypted to our servers where it is stored securely. The data is transmitted from the phones when the user is connected to WiFi and stored locally until then; if a successful transmission happens the data is removed from the device. We note that we do not collect user emails or explicit personal identifiers. The apps display a unique ID at the end of the survey for users to be allowed to contact us to ask for their data deletion. The user receives no medical advice through the app. To foster reproducibility, we will release the code of our apps as open source.

3.2 Crowdsourced dataset

Helped by a large media campaign orchestrated by the University, we were able to crowdsource data from a large number of users. In particular, as of May 22, 2020, our dataset is composed of 4352 unique users collected from the web app and 2261 unique users collected from the Android app, comprising 4352 and 3634 samples respectively. Of these, 235 declared that they have tested positive for COVID-19, 64 in the web form and 171 in the Android app. Of the Android users, 691 users contributed more than one sample, i.e., they returned to the app after two days and reported their symptoms and sounds again.

Figure 3 (a) shows the most frequent symptom distributions for those who have tested positive for COVID-19, 64 in the web form and 171 in the Android app. Of the Android users, 691 users contributed more than one sample, i.e., they returned to the app after two days and reported their symptoms and sounds again.

Figure 2: User Distribution: (a) Top 10 countries, and (b) Age. None=“Prefer not to Say”, None=Country not available.

The statistics of the data distributions are described below. All numbers quoted in this paragraph are aggregates across all active platforms unless stated otherwise. Figure 2 (a) illustrates the country (as recorded from location sample) distribution. We note that many users opted not to record their location. The gender breakdown is 4525 Male, 2056 Female, 26 Prefer not to say, and six Others. Of all completed surveys, 6088 had no symptoms and 3898 ticked at least one. Figure 2 (b) shows the age distribution, which is skewed towards middle age.

Figure 3 (a) shows the most frequent symptom distributions for all the Android platform users; we do not know which users have or have had COVID-19 recently but we know that only a small fraction of these have tested positive (see statistics above). In this group, the most common single symptom reported is the dry cough, while the most common combination of symptoms is a cough and sore throat. Figure 3 (b) shows the most frequent symptoms of the users who declared they had tested positive for COVID-19. Interestingly, the most common single symptoms are wet and dry cough and the most common combination is lack of sense of smell and chest tightness. This is aligned with the COVID-19 symptom tracker data [21]. The fact that the cough is one of the most reported symptom for COVID-19 but is also a general symptom of so many other diseases provides

---

1available at www.covid-19-sounds.org
We report here the number of samples used in our analysis after work to use only coughs and breathing (and not the voice samples) from users who reported dry or wet cough as symptoms. In 141 cough and breathing samples, 54 of these samples were from users positive for COVID-19 (in the last 14 days or before that) resulting from users who had a history of asthma, had not tested positive for COVID-19, and had a cough; these gave us 20 samples.

We intend to release all our data openly; however, due to the sensitive nature (e.g. voice) our institution has advised us to release it with one-to-one legal agreements with other entities for research purposes. Our web page will include information about how to access the data after publication.

4 METHODS

For the analysis of the sounds we followed standard data processing and modeling practices from the audio and sound processing literature targeting medical applications [25]. Based on the moderate size of the dataset selected, and the increased importance of explainability given the public health implications of our work, feature-based machine learning and shallow classifiers were employed for the classification tasks. In this section, we describe the extracted features and the methodology we followed to train robust classification models, taking into account specific idiosyncrasies of our data (e.g., longitudinal mobile users and cross-validation). We tested classifiers such as Logistic Regression (LR), Gradient Boosting Trees and Support Vector Machines (SVMs): we report best results in the results section, specifying which classifier gave them. We evaluated an SVM classifier with a Radial Basis Function (RBF) kernel. We considered different values of the following hyper-parameters: regularization parameter $C$ and kernel coefficient $\gamma$. Figure 4 illustrates the data processing pipelines.

4.1 Feature extraction

The raw sound waveform recorded by the Android app and the Web app is resampled to 22kHz as it is a standard resampling value for audio tasks. We used librosa [20] as our audio processing library. From the resampled audio various features are extracted at the frame and sample level, covering frequency-based, structural, statistical and temporal attributes. A complete list is provided below:

- **Duration**: the total size of the recording after trimming the leading and trailing silence from the signal.
- **Tempo**: the beats of the signals by measuring the onset strength, commonly used in music information retrieval [11]. In our context we use it for its peak detection capabilities (also see next bullet).
- **Onsets**: basic peak (onset) detector which locates onset events by picking peaks in an onset strength envelope (the envelope is a smooth curve outlining its extreme points).

As a control group, our analysis uses three sets of users. The first set consists of users from countries where the virus was not prevalent at the time (up to around 2000 cases): we treat these as non-covid users. We selected Albania, Bulgaria, Cyprus, Greece, Jordan, Lebanon, Sri Lanka, Tunisia, and Vietnam. Specifically, we define non-covid users as those with a clean medical history, who had never smoked, had not tested positive for COVID-19, and did not report any symptoms. These users contributed 298 samples. The second set non-covid with cough consists of users who meet the same criteria as the non-covid users, but declared a cough as symptom; these provided 32 samples. Finally, asthma with cough are the users who had a history of asthma, had not tested positive for COVID-19, and had a cough; these gave us 20 samples.

Figure 3: The most frequent 15 symptom combinations obtained from Android.

Figure 4 illustrates the data processing pipelines.
We now detail the evaluation of our methodology to classify COVID-19 audio samples from healthy ones using audio features described in Section 3. We calculate the FFT on the envelope and obtain the frequency with the highest amplitude.

- **RMS Energy**: the root-mean-square of the magnitude of a short-time Fourier transform which provides the power of the signal.
- **Spectral Centroid**: the mean (centroid) extracted per frame of the magnitude spectrogram.
- **Roll-off Frequency**: the center frequency for a spectrogram bin so that at least 85% of the energy of the spectrum in this frame is contained in this bin and the bins below.
- **Zero-crossing**: the rate of sign-changes of the signal.
- **MFCC**: Mel-Frequency Cepstral Coefficients obtained from the short-term power spectrum, based on a linear cosine transform of the log power spectrum on a nonlinear Mel scale. MFCCs are amongst the most common features in audio processing [9]. We use the first 13 components.
- **Δ-MFCC**: the temporal differential (delta) of the MFCC [3].
- **Δ²-MFCC**: the differential of the delta of the MFCC (acceleration coefficients) [3].

For the spectral features that generate time series (RMS Energy, Spectral Centroid, Roll-off Frequency and all variants of MFCCs) we extract several statistical features in order to capture the distributions beyond the mean. A complete list is: mean, median, root-mean-square, maximum, minimum, 1st and 3rd quartile, interquartile range, standard deviation, skewness, and kurtosis. The final feature matrix consists of 477 dimensions for each modality (cough, breath) and is further reduced by Principal Components Analysis (PCA) retaining a portion of the initial explained variance. More details about the pre-processing are provided in Section 5.

5 EVALUATION

We now detail the evaluation of our methodology to classify COVID-19 audio samples from healthy ones using audio features described in Section 4. Given the large class imbalance, we worked on a subsample of the initial collected dataset (described in Section 3.3). Firstly, we indicate how we merged the data from the different modalities and partitioned the dataset for our experiments. Findings and results are discussed in the later part of the section.

5.1 Experimental setup

**Classification tasks.** Based on the data collection (Section 3) we focus on three clinically meaningful binary classification tasks:

- **Task 1**: Distinguish users who have declared they tested **positive for COVID-19** (covid-tested) from users who have **not declared to have tested positive for COVID-19**, have a **clean medical history, never smoked**, are from **countries** where at the time COVID-19 was not prevalent and have a cough as a symptom (non-covid with cough).
- **Task 2**: Distinguish users who have declared they tested **positive for COVID-19** and have declared a cough as symptom (a prevalent symptom for COVID-19 tested users, as reported in Figure 3), (covid-tested with cough) from users who have **not declared to have tested positive for COVID-19**, have a **clean medical history, never smoked**, are from **countries** where at the time COVID-19 was not prevalent and have a cough as a symptom (non-covid with cough).
- **Task 3**: Distinguish users who have declared they tested **positive for COVID-19** and have declared a cough as symptom (covid-tested with cough), from users who have **not declared to have tested positive for COVID-19**, are from **countries** where at the time COVID-19 was not prevalent, have reported **asthma** as medical history and have a cough as a symptom (non-covid with cough).

**Data exploration.** As a first step after feature extraction, we examine the differences between the distributions of the cough features broken down by the respective class. Given the high dimensionality of the features, we cannot present all distributions, therefore we focus only on the **mean** statistical feature of each feature family (e.g., Centroid is Centroid **mean** here). For Task 1 (covid-tested/non-covid), the boxplots in Figure 5 show that coughs from covid-tested users are longer in total duration, have higher tempo, more onsets, higher period frequency, lower RMS, while their MFCC features [1st component and deltas] have fewer outliers. Similar trends are observed when we focus only on samples with reported cough symptoms (Task 2). Across both tasks, the samples from covid-tested users concentrate more towards the mean of the distributions, whereas the general (healthy) population shows greater span (inter-quartile range), with the hypothesis being that a (possibly forced) healthy cough is very diverse.

**Feature ablation studies.** In order to identify which audio modality (cough or breathing) contributes more to the classification performance, we repeat our experiments with three different audio inputs: **only cough**, **only breathing**, and **combined**. To account for the increasing dimensionality of the combined representation and to make for a fair comparison, we perform experiments to find the best cut-off value for PCA (see results in next section). The values of explained variance range between [70%, 80%, 90% and 95%]. In practice, this means that the with lower explained variance the classifiers will use fewer features and vice versa. Intuitively, a combined representation might need a more compressed representation than a representation using only coughs or breaths, to prevent overfitting.

**User based cross-validation.** We create training and test sets from disjoint user splits, making sure that samples from the same user do not appear in both splits. Note that this does not result in perfectly balanced class splits; however, we downsamples the majority (non-covid) class when needed. The test set is kept balanced.

Even then, it is not easy to guarantee that a split selects a representative test-set, so we performed a 10-fold-like cross validation using 10 different random seeds to pick disjoint users in the outer loop (80%/20% split), and a hyper-parameter search as inner loop to find the optimal parameters (using the 80% train-set in a 5-fold cross validation). Essentially, this setup resembles a nested cross-validation [7]. We conduct extensive experimentation by testing 1800 models (3 tasks x 3 modalities x 10 user splits x 4 dimensionality reduction cut-offs) using 5 hyper-parameter cross-validation runs). We selected several standard evaluation metrics such as the Receiver Operating Characteristic - Area Under Curve (ROC-AUC), Precision, and Recall. We report the average performance of the outer
Record cough and breathing sounds with a phone or a website.

1. Extract audio features such as MFCC etc. and populate a feature vector.
2. Train models to detect COVID sounds and evaluate their accuracy on the test set.
3. Split participants in disjoint 80/20 train and test sets.
4. PREPARE records and evaluate them on the test set.

Figure 4: Description of our machine learning pipeline, describing sounds input (coughs and breathing), the extracted feature vector, and our training and testing split of the users that are used to train classification models.

Figure 5: Boxplots of the mean features for Task 1.

Folds (10 user-splits) and the standard deviation. In the following section we report the performance of our three tasks.

**Confounders.** In order to make sure that exogenous information such as demographics does not confound the results, we control for age and sex by including them as one-hot-encoded features in our models (e.g. age group: 40-49 years old) and noticed that they do not improve or worsen the results substantially ($< \pm 2$ AUC). This suggests that the extracted features are invariant to demographics.

### 5.2 Distinguishing COVID-19 users from healthy users using coughs

Table 1 (first row) reports our classification results using just cough sounds analysis over Task 1 (as described above): the binary classification task of discriminating users who declare having tested positive for COVID-19 (covid-tested), against users who answered no to that question (non-covid). The metrics reported show that there seem to be some discriminatory signals in the data indicating that user coughs could be a good predictor when screening for COVID-19. In particular, the AUC for this task is at 72% while precision and recall are slightly short of 70%. Compared to the other tasks (Task 2 and 3), this task has the lowest standard deviations across the user-splits, mostly due to the bigger data size. We note that we applied a very simple classifier (Logistic Regression) and that the data is perhaps too limited in size to obviate the noise and diversity introduced by our crowdsourced data gathering (e.g., differences in microphones, surrounding noises, ways of inputting the sounds). Nevertheless these results give us confidence in the power of this signal.

### 5.3 Distinguishing COVID-19 coughs from other coughs

The second row of Table 1 describes the binary classification of users who reported they tested positive for COVID-19 and also declared a cough in the symptom questionnaire (a prevalent symptom for COVID-19 tested users, as reported in Figure 3) and a similar number of users who said they did not test positive for COVID-19 but declared a cough as symptom (Task 2). The results on all metrics show an AUC of 75%. Our precision for this task is at 82% showing
that our signal is able to distinguish quite well if a user has tested positive for COVID. However the recall is low, meaning that this model casts a small but very specialized net, does not detect a lot of COVID coughs, but there are almost only COVID coughs in the net. Nevertheless, the size of the data as well as the relatively high standard deviations compared to Task 1, render this result preliminary.

To reassure ourselves of the finding, we also compared the COVID-19 with cough users described above with users who said they did not test positive for COVID-19 but reported asthma in their medical history and declared a cough as a symptom. The results show an AUC of 76%. While the recall is acceptable, precision for this task is relatively low, likely due to the limited dataset for this task. However, this is a promising first result. We have further evaluated the utility of data augmentation for Task 2 and 3 to improve performance (Section 5.6).

### 5.4 Impact of breathing

Apart from using cough sounds as input, we tried to make use of breathing samples in combination with cough to improve classification. Figure 6a shows that for Task 1, breathing alone (at least for the simple features used) performs poorly (AUC around 60%); however in combination with cough sounds, it achieves the highest AUC and lowest standard deviation for the task (however, just marginally better than using coughs alone). The dimensionality size is not highly significant here, however, breathing seems to improve with more features.

For Task 2, in Figure 6b we observe the same trend of breathing alone not contributing much to the performance. Although the combined feature set achieves better AUC with lower dimensionality (PCA 80%), the cough modality seems to improve by using more features. This is expected, due to the different feature sizes. However, the high standard deviations do not show clear winners in terms of modalities. Lastly, Task 3 (Figure 6c) follows similar trends with overfitting becoming more apparent in higher dimensions, due to smaller sample size. Here, the cough modality outperforms the other modalities, and its performance increases up to a peak with PCA 90%. Overall, breathing sounds are promising, but only in combination with cough, which is the most informative modality.

### 5.5 Feature importance

Given the high feature dimensionality and potential computational cost to calculate them, it is reasonable to wonder which contribute most to classification. Here we examine this question. We used a range of classifiers, and therefore cannot compare different native feature importance methods because e.g., logistic regression provides interpretable coefficients but SVM with an RBF kernel transforms the features to a high-dimensional space; this transformation cannot be easily retrieved since it is implicit [15]. Moreover, the PCA features are not easily interpretable. Therefore, we remove a feature set one at a time (e.g., for a timeseries feature set we calculate many statistical attributes) and measure how much the overall AUC deteriorates, compared to using all features.

In Figure 7 we illustrate the feature importance for Tasks 1 and 2. For both tasks, the most important features come from the MFCC. However, the ranking changes for the rest of the features, for example $\Delta^2$-MFCC ranks 3rd for Task 1 but last for Task 2. Other important features for both tasks are the Tempo, $\Delta$-MFCC and the Onsets. The high importance of the $\Delta^X$-MFCC features suggests that the temporal dynamics of the coughs are significant (e.g. silences between the coughs or duration of cough bursts in relation to the next cough) and paves the way for time-aware models such as Recurrent Neural Networks in future work.

### 5.6 Data augmentation

To counter the small amount of control data available for our analysis in Tasks 2 and 3, we augmented the negative class (non-covid) for these two tasks by using three standard audio augmentation methods [27]: amplifying the original signal (1.15 to 2 times, picked using a random number), adding white noise (without excessively impacting signal to noise ratio), and changing pitch and speed of the original signal (0.8 to 0.99 times). We made sure not to distort the original signal significantly: we manually inspected and listened to the audio before and after performing the data augmentation. We applied each method twice to the original samples to obtain six times the number of original samples. Specifically, we increased the number of samples for 'non-covid with cough' and 'non-covid asthma cough'. Note that we used augmented samples only for training (the test set was kept intact). The results are shown in Table 2. We observe that the performance for all the metrics improved, particularly the recall and overall standard deviation, when compared to results in Table 1. With the much-improved recall our model is able to recognize a wide array of coughs, most importantly including almost all the COVID coughs. This is clinically important, since our aim is to identify COVID-19 positive cases; misclassifying

| Task                                      | Samples (users)* | Metric ± std |
|-------------------------------------------|------------------|--------------|
|                                           | ROC-AUC | Precision | Recall  |
| 1. covid-tested / non-covid               | 141 (62) / 298 (220) | 0.72 (0.05) | 0.67 (0.05) | 0.68 (0.09) |
| 2. covid-tested with *cough* / non-covid  | 54 (23) / 32 (29) | 0.75 (0.19) | 0.82 (0.20) | 0.54 (0.23) |
| 3. covid-tested with *cough* / non-covid  | 54 (23) / 20 (18) | 0.76 (0.30) | 0.64 (0.29) | 0.72 (0.31) |

**Table 1:** Classification results for the three tasks we evaluate using as input the cough sounds only. The number of samples before splitting to train/test and downsampling. Logistic Regression results are reported for the first task, while SVMs for the latter two tasks. We report the best representation size for PCA (detailed results for every cutoff are provided in Figure 6). Task 1 and Task 2: PCA = 0.95, Task 3: PCA = 0.8.
We have presented an ongoing effort to crowdsource respiratory sounds and study how such data may aid COVID-19 diagnosis. These results clearly only scratch the surface of the potential of this type of data; while our results are a positive signal, they are not as solid as necessary to comprise a standalone screening tool. We have, for the moment, limited ourselves to the use of a subset of all the data collected, to manage the fact that the proportion of COVID-19 positive reported users is considerably smaller than the rest of the users. We also have no ground truth regarding health status, and so we took users from countries where COVID-19 was not prevalent at the time as likely to be truly healthy when self-reporting as such (however this limited our dataset further). We are in the process of collecting more data through our software and discussing how this crowdsourced endeavor could be complemented by a controlled one, where we deliberately collect only COVID-19 tested user samples to use as ground truth. This will allow to analysis of a larger dataset, possibly with more advanced machine learning (e.g., deep learning). We are extending our study to voice features, which we have already collected from the users. Voice, as well as breathing and cough patterns, could give useful additional features for classification.

While we have preliminarily investigated the difference between COVID-19 coughs and asthma, our data records also users with other respiratory pathologies, and we hope to study this further to investigate how distinguishable COVID-19 is in this respect.

The mobile app reminds users to provide samples every couple of days: as a consequence we have a number of users for whom we could study the progression of respiratory sounds in the context of the disease. This is very relevant for COVID-19, and something we have not yet investigated in the current work.

ACKNOWLEDGMENTS

This work for this paper was supported by the European Research Council Advanced Grant EAR (Project 833296).

REFERENCES

[1] [n.d.]. CoughVid. https://coughvid.epfl.ch/about/, note= Accessed: 2020-05-30.
[2] [n.d.]. Detect Now: https://detectnow.org/, note= Accessed: 2020-05-30.
[3] [n.d.]. librosa.feature.delta. https://librosa.github.io/librosa/generated/librosa.feature.delta.html, note= Accessed: 2020-05-30.
[4] Charles Bales, Muhammad Nabeel, Charles N. John, Usama Masood, Haneya N. Qureshi, Hasan Farooq, Iryna Posokhova, and Ali Imran. 2020. Can Machine Learning Be Used to Recognize and Diagnose Coughs? arXiv:2004.01495 [eess.AS] 10 pages.
[5] Debrup Banerjee, Kari Islam, Keyi Xue, Gang Mei, Lemin Xiao, Guangfan Zhang, Roger Xu, Cai Lei, Shuwen Ji, and Jiang Li. 2019. A deep transfer learning approach for improved post-traumatic stress disorder diagnosis. Knowledge and Information Systems 60, 3 (2019), 1693–1724.
[6] I Brabenec, J Meňúška, Z Galaz, and Irena Rektorová. 2017. Speech disorders in Parkinson’s disease: Early diagnostics and effects of medication and brain stimulation. Journal of Neural Transmission 124, 3 (2017), 383–334.
[7] Gavan C Cawley and Nicola LC Talbot. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research 11, Jul (2010), 2079–2107.
[8] Yohan Chon, Nicholas D. Lane, Fan Li, Hojung Cha, and Feng Zhao. 2012. Automatically Characterizing Places with Opportunistic Crowdsensing Using Smartphones (UbiComp âĂŹ12). Association for Computing Machinery, New York, NY, USA, 481âĂŞ490. https://doi.org/10.1145/2370216.2370268

[9] Steven Davis and Paul Merollestein. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE transactions on acoustics, speech, and signal processing 28, 4 (1980), 357âĂŞ366.

[10] Gauri Deshpande and Björn Schuller. 2020. An Overview on Audio, Signal, Speech, & Language Processing for COVID-19. arXiv preprint arXiv:2005.08579 (2020).

[11] Daniel PW Ellis. 2007. Beat tracking by dynamic programming. Journal of New Music Research 36, 1 (2007), 51–60.

[12] Betul Erdogdu Sakar, Gorkem Serbes, and C. Okan Sakar. 2017. Analyzing the effectiveness of vocal features in early telediagnosis of Parkinson’s disease. PloS One 12, 8 (Aug. 2017), 1–18. https://doi.org/10.1371/journal.pone.0182428

[13] Maria Faurholt-Jepsen, Jonas Busk, Mads Frost, Maj Vinberg, Ellen M Christensen, Ole Winther, Jakob Eyvind Bardram, and Lars V Kessing. 2016. Voice analysis as an objective state marker in bipolar disorder. Translational Psychiatry 6, 7 (2016), e856. 8 pages.

[14] Jing Han, Kun Qian, Meishu Song, Zijiang Yang, Zhao Ren, Shuo Liu, Juan Liu, Huaiyuan Zheng, Wei Ji, Tomoya Kose. Xiao Li, Zixing Zhang, Yoshiharu Yamamoto, and Bjarne W. Schuller. 2020. An Early Study on Intelligent Analysis of Speech under COVID-19: Severity, Sleep Quality, Fatigue, and Anxiety. arXiv:2005.00096 [eess.AS] 5 pages.

[15] Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. 2008. Kernel methods in machine learning. The annals of statistics (2008), 1171–1220.

[16] Yinghui Huang, Sijun Meng, Yi Zhang, et al. 2020. The respiratory sound features of COVID-19 patients fill gaps between clinical data and screening methods. medRxiv (2020). https://doi.org/10.1101/2020.04.07.20051060 12 pages.

[17] Shih-Hong Li, Bor-Shing Lin, Chen-Han Tsai, Cheng-Ta Yang, and Bor-Shyh Lin. 2017. Design of Wearable Breathing Sound Monitoring System for Real-Time Wheeze Detection. Sensors 17, 1 (Jan. 2017), 1–15.

[18] Rajalakshmi Nandakumar, Shyamnath Gollakota, and Nathaniel Watson. 2015. Contactless sleep apnea detection on smartphones (MobiSys âĂŹ15). Association for Computing Machinery, New York, NY, USA, 45AÂŞ57. https://doi.org/10.1145/2742647.2742674

[19] Siyu Peng, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In Proceedings of the 14th python in science conference, Vol. 8.

[20] Jing Han, Kun Qian, Meishu Song, Zijiang Yang, Zhao Ren, Shuo Liu, Juan Liu, Huaiyuan Zheng, Wei Ji, Tomoya Kose. Xiao Li, Zixing Zhang, Yoshiharu Yamamoto, and Bjarne W. Schuller. 2020. An Early Study on Intelligent Analysis of Speech under COVID-19: Severity, Sleep Quality, Fatigue, and Anxiety. arXiv:2005.00096 [eess.AS] 5 pages.

[21] Jan Schlüter and Thomas Grill. 2015. Exploring Data Augmentation for Improved Singing Voice Detection with Neural Networks.. In ISMIR. 121–126.

[22] Neeraj Sharma, Prashant Krishnan, Rohit Kumar, Shreyas Ramoji, Prasanta Kumar Ghosh, and Sriiram Ganapathy. 2020. Coswara – A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis. arXiv:2005.10548 [eess.AS] 5 pages.