Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Pay Attention to the Speech: COVID-19 Diagnosis using Machine Learning and Crowdsourced Respiratory and Speech Recordings

Mahmoud Aly, Kamel H. Rahouma, Safwat M. Ramzy

PII: S1110-0168(21)00585-8
DOI: https://doi.org/10.1016/j.aej.2021.08.070
Reference: AEJ 2572

To appear in: Alexandria Engineering Journal

Received Date: 3 June 2021
Revised Date: 7 August 2021
Accepted Date: 23 August 2021

Please cite this article as: M. Aly, K.H. Rahouma, S.M. Ramzy, Pay Attention to the Speech: COVID-19 Diagnosis using Machine Learning and Crowdsourced Respiratory and Speech Recordings, Alexandria Engineering Journal (2021), doi: https://doi.org/10.1016/j.aej.2021.08.070

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2021 Production and hosting by Elsevier B.V. on behalf of Faculty of Engineering, Alexandria University.
Pay Attention to the Speech: COVID-19 Diagnosis using Machine Learning and Crowdsourced Respiratory and Speech Recordings
Mahmoud Aly, Kamel H. Rahouma, Safwat M. Ramzy

aDepartment of Electrical Engineering, Faculty of Engineering, Minia University, Minia, Egypt
bDepartment of Electrical Engineering, Faculty of Engineering, Sohag University, Sohag, Egypt

Corresponding author: Mahmoud Aly
Mailing address: Department of Electrical Engineering, Faculty of Engineering, Minia University, 61519, Minia, Egypt
Mobile phone: 201553763638 - 201116984486
E-mail: engma7moud3ly@gmail.com

Abstract
Since the outbreak of COVID-19, many efforts have been made to utilize the respiratory sounds and coughs collected by smartphones for training Machine Learning models to classify and distinguish COVID-19 sounds from healthy ones. Embedding those models into mobile applications or Internet of things devices can make effective COVID-19 pre-screening tools afforded by anyone anywhere. Most of the previous researchers trained their classifiers with respiratory sounds such as breathing or coughs, and they achieved promising results. We claim that using special voice patterns besides other respiratory sounds can achieve better performance. In this study, we used the Coswara dataset

1 Corresponding author, Email: engma7moud3ly@gmail.com
where each user has recorded 9 different types of sounds as cough, breathing, and speech labeled with COVID-19 status. A combination of models trained on different sounds can diagnose COVID-19 more accurately than a single model trained on cough or breathing only. Our results show that using simple binary classifiers can achieve an AUC of 96.4% and an accuracy of 96% by averaging the predictions of multiple models trained and evaluated separately on different sound types. Finally, this study aims to draw attention to the importance of the human voice alongside other respiratory sounds for the sound-based COVID-19 diagnosis.

Keywords: COVID-19, Machine Learning, Cough sounds, Speech, Respiratory sounds

1. Introduction

COVID-19 (coronavirus disease 2019) is caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1]. It has been spreading around the world since December 2019. Until the beginning of May 2021, more than 149 million confirmed cases of COVID-19 were reported with more than 3 million deaths [2]. Public health practices such as mask-wearing and social distancing will continue to be important until a sufficient proportion of the population is immunized or achieves herd immunity [3]. The genetic material identification of the virus by RT-PCR is the gold standard test for COVID-19, but new diagnostic methods with higher sensitivity and specificity, as well as faster results, are necessary [4]. As RT-PCR tests are time-consuming, expensive, and not easily available all the time in all regions. Fast, low-cost, widely accessible solutions and prescreening tools are required to limit the spread of COVID-19 and forward the suspected people to the clinical examination. As dry cough is one of the most common symptoms of COVID-19 [5] and about 71.4% of 31845 confirmed cases presented with symptoms of cough and dyspnea [6], early efforts have been made to prove the possibility of using respiratory sounds, coughs, and speech in diagnosing COVID-19. In [7] speech and human audio analysis are found to be most useful for COVID-19 diagnosis. [8] found that audio-only-based diagnosis is possible for COVID-19 using Speech and respiratory sounds analysis and machine learning. Those studies were followed by many efforts aimed to detect and diagnose COVID-19 from the cough and respiratory sound recordings using Machine learning and Deep learning-based models, and our study contributes towards this approach with proven recommendations to use the human voices in COVID-19 diagnosis. In the next section, we will discuss some of the significant previous researches.

2. Related Works

Since the beginnings of the pandemic, many projects were released for the sake of collecting coughs and other respiratory sounds by websites and mobile applications to be used in future researches. Users are allowed to record, upload recordings, and attach other clinical information. As there were no standard rules for data collection and sound recording, the several datasets come with different audio formats, bit rates, and different types of recordings like breathing, coughs, speech, or all of those types together. Coswara [9] and Coughvid [10] are examples of publicly available crowdsourced datasets which size is growing day by day.
On the other hand, many promising efforts were made for COVID-19 sounds classification using Machine learning and Deep Learning-based models. In [11] Imran et al. (2020) started a preliminary study for COVID-19 diagnosis from cough samples via an app, they trained different variants of deep learning and classical machine learning algorithms, which could distinguish COVID-19 coughs and several types of non-COVID19 coughs. It has not been a long time since researchers from Cambridge University [12] established a COVID-19 sound collection framework [13] to collect coughs and breathing sounds. Data collected is used to train their models to distinguish different cough types to achieve AUC above 80% for all tasks. In [14] they claimed to collect the largest cough dataset with verified RT-PCR tests. They developed a CNN-based framework that directly predicts a binary classification label indicating the probability of the presence of COVID-19 with AUC 72%. Also, researchers at MIT built a data collection pipeline of COVID-19 cough recordings [15], and as discussed in [16] they used a Convolutional Neural Network (CNN) based architecture made up of one Poisson biomarker layer and 3 pre-trained ResNet50’s in parallel trained on a balanced dataset of 5320 samples to achieve an AUC of 97%. In [17] they designed a COVID-19 diagnosis AI framework based on the cough sounds and other symptoms metadata, they created a 30000 cough sample segmented out of 328 cough samples. By embedding other symptoms with coughs, they achieved an accuracy of 96.8%. Another study that used clinical features besides cough recordings is Virufy [18]. In Virufy they developed an AI-based ensemble model of 3 separate networks. They used both Coswara and Coughvid [9,10] datasets for training their models and collected their testing datasets to predict COVID-19 infection with an AUC of 77.1%. In [19] Pahar et al. (2020) used the Coswara dataset [9] and another dataset collected from south Africa to train several models that could discriminate between the COVID-19 positive and the healthy coughs to achieve an AUC of 98% with Resnet50 classifier and 94% with LSTM classifier. In [20] Coppock et al. (2021) used the same dataset of the Cambridge study [12], but they combined breath and coughs samples together, and used an end-to-end CNN-based model with Resnet architecture to achieve an AUC of 84.6% in symptomatic and asymptomatic COVID-19 recognition.

In this study, we used the Coswara dataset [9] to train our models because it has various types of sounds such as breathing, coughs, and other special voice patterns. For the testing purpose, we used two datasets described in the Virufy study [18] besides samples from the Coswara dataset. Our models didn’t rely on clinical symptoms like [17,18] but only cough and respiratory sounds. In Table 1 there is a comparison to many previous works and their methodologies and results.

### Table 1: Comparison to many previous studies for the sound-based COVID-19 diagnosis.

| Research | Dataset | Sound type | Features | Models/Classifiers | Results (AUC) |
|----------|---------|------------|----------|--------------------|---------------|
| Fakhry et al. [21] (Virufy) | Coughvid [10] | Cough | Mel-spectrogram, MFCC, Clinical features | Multi-Branch Deep Learning Network | micro-average AUC of 91% |
| Coppock et al. [20] | Covid-19-sounds [13] | Cough, Breathing | Mel-spectrogram | ResNet | 84.6% |
| Pahar et al. [19] (University of Stellenbosch) | Coswara [9], SARCOS [19] | Cough | MFCC - Log Energies - ZCR - Kurtosis... | ResNet50, LSTM | 98%, 94% |
| Chaudhari et al. [18] (Virufy) | Coswara [9], Coughvid [10], Virufy [18] | Cough | Mel-spectrogram, MFCC, Clinical features | Ensemble Deep Learning Model | 77.1% |
| J. Laguarta et al. [16] (MIT) | Opensigma [15] | Cough | MFCC - other biomarkers | ResNet50 | 97% |
3. Data Collection

3.1 Training Data

To train our models, we used the Coswara dataset [9]. The sound samples are collected via worldwide crowdsourcing using the internet. All samples are available publicly here \(^2\). The released dataset (until 19 April 2021) was used for training our models. The training data contained 1604 samples from different users. Samples are labeled with 7 COVID-19 statuses as shown in Table 2.

Table 2: Different COVID-19 cases in the Coswara dataset.

| COVID-19 status                  | Number of Samples |
|---------------------------------|-------------------|
| Healthy                        | 1249              |
| No_resp_illness_exposed         | 116               |
| Positive_mild                  | 96                |
| Resp_illness_not_identified     | 84                |
| Recovered_full                 | 28                |
| Positive_asympt                | 18                |
| Positive_moderate              | 13                |
| Total                          | 1604              |

We selected all healthy cases 1249 and labeled them as Negative. On the other side, we selected the recordings labeled with (positive_mild, positive_asympt, positive_moderate) 127 and labeled them as Positive. Every participant in the dataset has 9 different sound types (Breathing-deep, Breathing-shallow, Cough-heavy, Cough-shallow, Counting-fast, Counting-normal, Vowel-A, Vowel-E, Vowel-O), the data is described in detail here [9]. There were missing sounds for some users, so we selected the users who recorded all the 9 sound types. The final results were 9 separate datasets; each dataset contains samples from a single sound type labeled with a COVID-19 status (positive) or (negative). The total samples in each dataset were 1299, where (126) positives & (1173) negatives.

\(^2\) https://github.com/iiscleap/Coswara-Data
The main reason for selecting the Coswara dataset rather than other available datasets, is the variety of respiratory sounds like breathing, coughs, and different human voices, so it serves our study a lot to confirm the importance of different respiratory sounds as well as speech in achieving more accurate diagnosis for COVID-19.

3.2 Testing Data

For testing our models, we used samples from 3 different publicly available datasets. The first testing samples were selected from the same dataset used for the training; Coswara [9], but we used the recordings released after the date (19 April 2021). Selecting (healthy) labeled samples, as well as all (positive_mild, positive_asympt, positive_moderate), produced 9 datasets of different sound types, each dataset contains 10 positive samples versus 14 negative samples.

Two additional datasets were used. The first is the Virufy (Open Cough dataset³), and the second is the Virufy (Clinical-India-1) dataset⁴. Although the datasets available publicly are not the same as they described in their research [18]. But we can use them anyway to evaluate our models' performance against sounds from different sources. The testing datasets are described in Table 3.

| Dataset Name               | Sound Types               | Negative Samples | Positive Samples | Total  |
|----------------------------|---------------------------|------------------|------------------|--------|
| Coswara                    | Cough - Breathing - Speech| 14               | 10               | 24     |
| Virufy Open Dataset        | Cough                     | 9                | 7                | 16     |
| Virufy Clinical 1          | Cough                     | 7                | 60               | 67     |

3.3 Feature Extraction

For analyzing sounds and extracting features we used Librosa: “an audio signal analysis framework in Python” [22]. The raw audio samples are resampled to 22.50 kHz and processed without segmentation as in [16,18]. The features are extracted for every sound type such as (breathing, cough, speech) in a separate dataset to produce 9 different datasets of features. Like most previous studies, we used the Mel Frequency Cepstral coefficients (MFCC) features [12,14,16,21]. Also using Librosa we extracted other 5 important features that were (Spectral Bandwidth, Spectral Centroid, Spectral Roll-off, Zero Crossing Rate, RMS Energy). Besides those features, other 4 statistical methods (Skewness, Kurtosis, Variation, SEM) were applied to the MFCCs to produce more features that were very effective on the models’ performance. The last feature is the (Muscle Fatigue) biomarker introduced by MIT [16] which had a significant improvement on the classifiers’ performance. In the following points, we describe the extracted features in more detail.

- Mel Frequency Cepstral coefficients (MFCCs)

---
³ https://github.com/virufy/virufy-data
⁴ https://github.com/virufy/virufy-cdf-clinical-india-1
Mel Frequency Cepstral coefficients (MFCCs) represent the short-term power spectrum of a signal on the Mel-scale of the frequency [17]. To extract the MFCCs, first, the Discrete Fourier Transform (DFT) is applied to an audio sample as follows:

\[ Y(k) = \sum_{t=0}^{N-1} y_t[t] w(t) \exp \left( -\frac{2\pi ik}{N} \right), k = 0,1,...,N - 1 \]  

(1)

Where \( N \) is the frame samples. \( y_t[t] \) is the audio signal in the discrete-time domain. \( w(t) \) is the window function in the time domain. \( Y(k) \) is the \( k^{th} \) harmonic corresponding to the frequency \( f(k) = kF_s/N \) where \( F_s \) is the sampling frequency.

Then a triangular bandpass filter is used with the DFT output. Finally, the Discrete Cosine Transform (DCT) is applied to the output of the filters to get the MFCCs as follows:

\[ MFCC(i) = \sqrt{\frac{2}{M}} \sum_{m=1}^{M} \log(E(m)) \cos \left( \frac{\pi i}{M} (m - 0.5) \right), i = 1,2,...,I \]  

(2)

Where \( i \) is the coefficient order, \( I \) is the number of coefficients. \( E(m) \) is the bandpass filter energies. \( M \) is the total number of bandpass filters. Noteworthy, that we selected the first 39 MFCCs for every audio sample.

- **Spectral Centroid**

  The spectral centroid \( C \) is the barycenter of the spectrum [23].

  \[ C = \frac{\sum_{k=b_1}^{b_2} f_k S_k}{\sum_{k=b_1}^{b_2} S_k} \]  

(3)

Where \( f_k \) is the frequency in Hz corresponding to bin \( k \). \( S_k \) is the spectral value at bin \( k \). \( b_1 \) and \( b_2 \) are the band edges in bins over which to calculate the spectral centroid [24].

- **Spectral Bandwidth**

  The bandwidth of the spectrum is described by spectral spread [25]. The spectral bandwidth is the spread of the spectrum around its mean value [23] it can be described as.

  \[ spread = \sqrt{\frac{\sum_{k=b_1}^{b_2} (f_k - \mu_1)^2 S_k}{\sum_{k=b_1}^{b_2} S_k}} \]  

(4)

Where \( f_k \) is the frequency corresponding to bin \( k \). \( S_k \) is the spectral value at bin \( k \). \( b_1 \) and \( b_2 \) are the band edges, in bins, over which to calculate the spectral centroid. \( \mu_1 \) is the spectral centroid described in equation (3) [26].

- **Spectral Roll-off**
The spectral roll-off point is the value of the frequency so that 95% of the signal energy is contained below this frequency [23] and it is represented as follows:

$$\sum_{k=b_1}^{i} s_k = k \sum_{k=b_1}^{b_2} s_k$$

(5)

Where $S_k$ is the spectral value at bin $k$. $b_1$ and $b_2$ are the band edges in bins over which to calculate the spectral spread. $k$ is the percentage of the total energy contained between $b_1$ and $i$, it can be 95%, 85%, etc [27].

- **Zero-Crossing Rate (ZCR)**

Zero-Crossing Rate is the number of times the signal changes its sign within a frame [19].

$$ZCR = \frac{1}{T} \sum_{t=1}^{T-1} \lambda(S_t S_{t-1} < 0)$$

(6)

Where $\lambda = 1$ when the sign of $S_t$ and $S_{t-1}$ differ, and $\lambda = 0$ when the sign of $S_t$ is the same as $S_{t-1}$.

- **RMS Energy**

RMS energy is the root mean square of the magnitude of a short-time Fourier transform which provides the power of the signal [12, 25].

$$RMS = \sqrt{\frac{1}{N} \sum_{i=0}^{N-1} x(n+1)^2}$$

(7)

Where $n$ is a discrete-time index, and $N$ is the size of the analysis frame.

- **Skewness**

Skewness is the third-order moment of a signal, which measures the symmetry in a probability distribution [17].

$$Skewness = \frac{E(\gamma[t] - \mu)^3}{\sigma^3}$$

(8)

Where $\mu$ is the mean and $\sigma$ is the standard deviation of $\gamma[t]$.

- **Kurtosis**

Kurtosis is the fourth-order moment of a signal, which measures the peakiness or heaviness associated with the audio probability distribution [17].

$$Kurtosis = \frac{E(\gamma[t] - \mu)^4}{\sigma^4}$$

(9)

Where $\mu$ is the mean and $\sigma$ is the standard deviation for $\gamma[t]$.

- **Coefficient of Variation**
The coefficient of variation is the ratio of the biased standard deviation to the mean [28].

\[ Variation = \frac{s}{x} \]  

(10)

Where \( s \) is the standard deviation, and \( x \) is the arithmetic mean [29].

- **SEM**

SEM is the standard error of the mean [29].

\[ SEM = \frac{s}{\sqrt{n}} \]  

(11)

Where \( s \) is the standard deviation, and \( n \) is the number of values.

Finally, all the sound types produced 9 separate datasets of features (Breathing-deep, Breathing-shallow, Cough-heavy, Cough-shallow, Counting-fast, Counting-normal, Vowel-A, Vowel-E, Vowel-O). Every dataset has the shape of 1299x50, where 1299 is the total number of samples in each dataset, and 50 contains the 49 extracted features plus the label. All the extracted features are summarized in Fig.1.

![Fig. 1: The extracted features from one audio sample](image)

### 3.4 Dataset Balancing

The lack of positive samples was the most frequent problem that meets who wants to classify the respiratory sounds. In the Coswara dataset [9] as described in section 3.1, the ratio of negative samples to positives is nearly 10:1. This is called a deservedly imbalanced dataset, so we worked on balancing the datasets by oversampling.

For the effectiveness of the proposed method in [19], the oversampling technique (SMOTE) is applied to balance the minor samples (positive) to the major samples (negative) by generating synthetic positive samples using the Imbalanced-learn library [30]. At first, the features described in Section 3.3 were standardized by removing the mean and scaled to unit variance by using a standard scaler algorithm introduced by the Sklearn library [31]. Then only the
training dataset is oversampled to make a class-balanced dataset with a positive to negative ratio of 1:1 as indicated in Fig 2. It is noteworthy that we didn’t oversample the testing data.

![Fig. 2: The number of samples in the training dataset (a) The actual data. (b) Balanced data after applying oversampling to the minor positive class.](image)

4. Classification

4.1 Model

The network architecture of our deep learning model is very simple as shown in Fig. 3. It accepts inputs with a shape of 49 as the number of features. Followed by a single hidden dense layer with 16 nodes and a ReLU activation function. The dropout layer at a rate of 0.5 is used to reduce the overfitting. Finally, we see the output layer with a single node and a sigmoid activation function. The model is trained using a Binary Cross Entropy as loss function, Adam optimizer, and a learning rate of 0.001 as used in [18]. The classification process was implemented using the Tensorflow framework [32].

![Fig. 3: Model architecture.](image)

4.2 Training Strategy

We trained our models using the training datasets mentioned in Section 3.1. Each sound type has a separate dataset of 1299 samples in total, with 1173 negative samples and 126 positives. The training data is oversampled as described
in Section 3.4 to balance the negative samples with the positives, to produce a balanced dataset with a total of 2346 samples for each sound type. Out of the total samples, 20% are separated for validating the models while the rest are used for training the models. With the same architecture and parameters mentioned in Section 4.1, we trained 9 separate models with 9 different datasets (Breathing-deep, Breathing-shallow, Cough-heavy, Cough-shallow, Counting-fast, Counting-normal, Vowel-A, Vowel-E, Vowel-O). Sounds of different types like (breathing and cough) can’t be used to train a single model unlike what was introduced in [20]. We preferred to train and classify every sound type with a separate classifier, and the performance of each classifier is evaluated individually. Finally, we average the predictions of several classifiers tested on relevant sound types, and the average of predictions can be used to measure the accuracy and other metrics for our classifiers’ performance as shown in Fig. 4.

![Fig. 4: The strategy of models training and evaluation with different types of COVID-19 sounds.](image)

### 4.3 Single Classifier

We started to train the models individually, each model is trained with a unique sound type to see which type has more information about COVID-19, and which types can’t be used for the diagnosis. To evaluate the performance, the models were tested with the test samples mentioned in Section 3.2 taken from the Coswara dataset [9].

#### 4.3.1 Breathing Sounds

The model tested with the shallow breathing samples achieved an AUC of 0.81, while the model tested with the deep breathing samples had a small AUC compared to the other, as shown in Table 4. The confusion matrix for every classifier is shown in Fig. 5. The shallow breathing model could correctly detect 6 positive samples (TP) out of 10, and 12 negative samples (TN) out of 14 as shown in Fig. 5 b. Although the accuracy of the Deep Breathing classifier
is small, we can’t generalize that the deep breaths can never be used to diagnose COVID-19, because both of the training and testing data are small, and the breaths might not be recorded as they should be. But we can confirm that respiratory sounds, especially shallow breaths contain the signature of COVID-19.

### Table 4: The classification results for breathing sounds.

| Metric | Deep Breathing | Shallow Breathing |
|--------|----------------|------------------|
| Accuracy | 0.58 | 0.75 |
| AUC | 0.36 | 0.811 |

![Confusion Matrix](image)

**Fig. 5**: The confusion matrix for (a) Deep Breathing classifier. (b) Shallow Breathing classifier.

#### 4.3.2 Cough Sounds

In the second experiment, both the heavy cough and the shallow cough models detected the same number of COVID-19 samples, but the overall performance of the heavy cough model was better than the shallow cough model to achieve an AUC of 0.84 as shown in Table 5. The confusion matrix for every cough classifier is shown in Fig. 6. In Fig.6 (a) we see that the heavy cough model could correctly detect 5 positive samples (TP) out of 10 samples, and 13 negative samples (TN) out of 14. While the shallow cough model in Fig.6-b detected correctly 5 positive samples (TP), and 10 negative samples (TN).

### Table 5: The classification results for cough sounds.

| Metric | Heavy Cough | Shallow Cough |
|--------|-------------|---------------|
| Accuracy | 0.75 | 0.625 |
| AUC | 0.84 | 0.732 |
4.3.3 Speech (counting)

The classification results for the fast counting and normal counting classifiers didn’t show promising performances as shown in Table 6 and indicated in Fig. 7. But both models are still useful in multi-models classification as we will see later.

Table 6: The classification results for counting sounds.

| Metric       | Normal Counting | Fast Counting |
|--------------|-----------------|---------------|
| Accuracy     | 0.54            | 0.58          |
| AUC          | 0.55            | 0.56          |
4.3.4 Speech (Vowels)

Other human voices like the vowels could have a classification performance better than counting sounds. As shown in Table 7, the vowels E and O achieved an accuracy of 0.71 which is higher than the accuracy of the vowel A that is 0.54. The confusion matrix for every classifier is shown in Fig. 8. We can see that the vowel E classifier in Fig. 8 (b) could correctly detect 7 positive samples (TP) out of 10 which is the best COVID-19 detector till now.

Table 7: The classification results for Vowels A, E, and O.

| Metric | Vowel A | Vowel E | Vowel O |
|--------|---------|---------|---------|
| Accuracy | 0.54    | 0.71    | 0.71    |
| AUC    | 0.61    | 0.72    | 0.75    |

Fig. 7: The confusion matrix for (a) Normal Counting classifier. (b) Fast Counting classifier.

Fig. 8: The confusion matrix for (a) Vowel A classifier. (b) Vowel E classifier. (c) Vowel O classifier.
4.4 Multiple Classifiers

From the previous section, it is clear that the best performance was achieved by the models trained and evaluated using the heavy cough and the shallow breathing datasets, which achieved AUCs of (0.84, 0.81) respectively. Although the speech model trained with vowel (E) dataset could detect the largest number of positive samples among all the models. So it is not wise to select a single model and a single sound type to build a sound-based COVID-19 detector that is alleged to be accurate and reliable. But we can use the predictions generated from two or more models and calculate the average of those predictions to achieve the best classification results. For this purpose, let’s assume that we have a set of 9 models (S) trained on 9 different datasets, so $S = \{\text{Breathing-deep, Breathing-shallow, Cough-heavy, Cough-shallow, Counting-fast, Counting-normal, Vowel-A, Vowel-E, Vowel-O}\}$. If we want to know the set of models that achieves the best-averaged predictions, we have to test all the possible subsets of models in the set (S). The number of all possible subsets in the set (S) is known as the Power Set $P(S)$. In mathematics, the power set is the set of all subsets including the empty set. Hence the number of subsets $P(S) = 2^9 = 512$. While an empty set means no models selected, so all the possible combinations of models are (511).

In table 8, we can see random combinations of models.

| Index | Model 0            | Model 1               | Model 2               | Model 3      |
|-------|--------------------|-----------------------|-----------------------|--------------|
| 1     | Breathing-deep     |                       |                       |              |
| 20    | Breathing-deep     | Counting-fast         |                       |              |
| 50    | Breathing-deep     | Breathing-shallow     | Counting-normal       |              |
| 121   | Breathing-deep     | Vowel-E               | Vowel-O               |              |
| 200   | Breathing-deep     | Breathing-shallow     | Cough-heavy           | Vowel-O      |
| 250   | Breathing-deep     | Vowel-A               | Vowel-E               | Vowel-O      |

We have to test the data on all the 511 sets of models, and take the average of the predictions for all the models in a single set to see which models set or combination achieves the best performance.

4.5 Statistical Prediction

As mentioned in section 4.1, in the output layer we used the sigmoid activation function which returns the predicted output $P$ in the range of 0 to 1. If the value of $P > 0.5$, then the sample is classified as (positive), otherwise it is classified as (negative), where the value 0.5 is the default classification threshold. Let’s assume a single confirmed positive COVID-9 case which is tested on the model set $\{\text{breathing-deep, Breathing-shallow, Counting-normal}\}$, and the predicted outputs for the models were (0.7, 0.4, 0.8) respectively. We see that the Breathing-shallow classifier made a misleading prediction (0.4) which is classified as (negative), while the other classifiers correctly classified the samples as (positive) with predictions $> 0.5$. So we can’t rely on a single model, but it is preferred to use one of the following statistical methods applied to the predictions of several models to find out the most realistic predictions.

- **Mean of the Predictions**
Here we calculate the mean of predictions of different models. In the last example for the values (0.7, 0.4, 0.8) that are predicted by the models set \{Breathing-deep, Breathing-shallow, Counting-normal\} the mean equals to (0.63) which is classified correctly as (positive) case, as shown in Fig. 9.

![Fig. 9](image)

**Fig. 9**: The mean of values predicted by 3 models set to a confirmed positive case of COVID-19.

- **Median of the Predictions**

As an alternative to the mean, we can use the median function to calculate the median value of the predictions. Consider the predicted values of the models set \{Breathing-deep, Breathing-shallow, Counting-normal\} were (0.7, 0.1, 0.51) respectively. Hence the mean of the values equal to (0.44) which is considered as (negative). While the median value is (0.51), so the case is correctly classified as (positive) as shown in Fig.10.

![Fig. 10](image)

**Fig. 10**: The median of values predicted by 3 models set to a confirmed positive case of COVID-19.

- **Maximum of the Predictions**

Assume a set of 4 models \{Breathing-deep, Breathing-shallow, Cough-heavy, Counting-normal\} predicted the values (0.55, 0.6, 0.4, 0.3) for a confirmed positive case, hence the mean and the median of the 4 values will be (0.46, 0.48) respectively, and both values are considered as (negative). So in this case, we use the maximum function rather than the mean or the median. Hence the max of (0.55, 0.6, 0.4, 0.3) is (0.6) which is considered as a (positive), and the case is classified correctly as shown in Fig. 11.
Fig. 11: The maximum of the values predicted by 4 models set to a confirmed positive case of COVID-19.

- **Smaller Classification Threshold**

Changing the classification threshold to a smaller value might help in some cases. In the last section, we found that the mean and the median of the 4 predicted values were (0.46, 0.48) respectively. So comparing those values to a classification threshold of the value (0.5) will classify the case as (negative). But if we changed the threshold to (0.4), then both the median and the mean will be classified correctly as (positive).

5. Evaluation

Here we explain the evaluation process for our models on the three testing datasets which are described in Section 3.2. For each dataset besides the accuracy, we measured the AUC as it was the most frequently used performance measure in the previous researches. Also, we clarified the classification performance for every dataset with the confusion matrix to determine the numbers of samples that were detected correctly.

5.1 Coswara Dataset

After testing the data with all models sets, a set of 6 models \{Breathing-shallow, Cough-heavy, Cough-shallow, Counting-fast, Vowel-E, Vowel-O\} achieved an accuracy of 95.8% and AUC of 96.4% by calculating the median of the predictions at a classification threshold (0.4), as shown in Table 9. Also, this set of models could detect all the positive cases as indicated in The confusion matrix shown in Fig. 12 (a).

Table 9: Evaluating Coswara testing dataset on different sets of models by calculating the median of predictions.

| Dataset Name | Classification Threshold | Statistical Method | Accuracy |
|--------------|--------------------------|--------------------|----------|
| Coswara      | 0.4                      | Median             |          |
| Models set   | Breathing-shallow, Cough-heavy, Cough-shallow, Counting-fast, Vowel-E, Vowel-O | 0.958              |
|              | Cough-heavy, Cough-shallow, Vowel-E | 0.917              |
|              | Vowel-E, Vowel-O         | 0.917              |

By the same logic on calculating the mean of the predictions, we found the models set \{Breathing-shallow, Cough-shallow, Vowel-O\} achieved an accuracy of 91.7% and AUC of 91.4% at the same classification threshold (0.4), as
shown in Table 10. In this case, the models detected correctly 9 positive cases out of 10 as indicated in the confusion matrix shown in Fig. 12 (b).

**Table 10:** Evaluating Coswara testing dataset on different sets of models by calculating the mean of predictions.

| Dataset Name | Classification Threshold | Statistical Method | Models set                                      | Accuracy |
|--------------|--------------------------|--------------------|-----------------------------------------------|----------|
| Coswara      | 0.4                      | Mean               | Breathing-shallow, Cough-shallow, Vowel-O     | 0.917    |
|              |                          |                    | Breathing-shallow, Cough-shallow, Counting-fast, Vowel-O | 0.917    |
|              |                          |                    | Breathing-shallow, Vowel-O                     | 0.875    |

Fig. 12: The confusion matrix for the classification of Coswara testing dataset (a) By calculating the median of the predictions of the models set (Breathing-shallow, Cough-heavy, Cough-shallow, Counting-fast, Vowel-E, Vowel-O). (b) By calculating the mean of the predictions of the models set (Breathing-shallow, Cough-shallow, Vowel-O).

5.2 Virufy Dataset

As mentioned in section 3.2, Both (Virufy open dataset) and (Virufy clinical 1 dataset) contain only cough recordings, hence as an exception, a single cough sample is tested on two or more different models trained on different sound types.

On testing the Virufy open dataset, the best performance is achieved by the models set \{Breathing-shallow, Cough-shallow\} with an Accuracy of 75% and a similar AUC by calculating the mean or the median of predictions at classification threshold equals to (0.35) as shown in Table 11. This set of models could detect 5 positive samples out of 7 as shown in Fig. 13 (a).
Table 11: Evaluating Virufy open dataset on different sets of models by calculating the (Median or Mean) of the predictions.

| Dataset Name      | Classification Threshold | Statistical Method | Accuracy |
|-------------------|--------------------------|--------------------|----------|
| Virufy open Dataset | 0.35                     | Mean | Median | 0.75 |
| Models set        |                          | Breathing-shallow, Cough-shallow | 0.75 |
|                   |                          | Breathing-deep, Breathing-shallow | 0.625 |
|                   |                          | Cough-shallow | 0.5 |

By the same logic, on testing the Virufy clinical 1 dataset, a set of 6 models \(\{\text{Breathing-deep, Breathing-shallow, Cough-heavy, Cough-shallow, Counting-normal, Vowel-E}\}\) could achieve an accuracy of 77.6% by calculating the maximum of predictions at classification threshold equals to (0.5), as shown in Table 12. This set of models could detect 49 positive samples out of 60 as shown in Fig. 13 (b).

Table 12: Evaluating Virufy clinical 1 dataset on different sets of models by calculating the maximum of the predictions.

| Dataset Name      | Classification Threshold | Statistical Method | Accuracy |
|-------------------|--------------------------|--------------------|----------|
| Virufy Clinical 1 | 0.5                      | Maximum            | 0.776    |
| Models set        |                          | Breathing-deep, Breathing-shallow, Cough-heavy, Cough-shallow, Counting-normal, Vowel-E | 0.776 |
|                   |                          | Breathing-deep, Breathing-shallow, Cough-heavy, Counting-normal, Vowel-E | 0.761 |

![Confusion matrix](image)

**Fig. 13**: The confusion matrix for the classification of (a) Virufy open Dataset by calculating the mean of the predictions of the models set (Breathing-shallow, Cough-shallow) at a classification threshold of 0.35. (b) Virufy Clinical 1 Dataset by calculating the maximum predicted values of the set of models (Breathing-deep, Breathing-shallow, Cough-heavy, Cough-shallow, Counting-normal, Vowel-E) at a classification threshold of 0.5.
6. Discussion and Summary of Contribution

Regarding the comparison between our results to other previous studies that used similar methods, features, and the same dataset, we claim that we have achieved very promising results. Again and again, we settle that using different sound types together (breathing, cough, and speech) is essential for building reliable and more accurate COVID-19 sound classifiers. The classification results for some previous studies that used the Coswara dataset are presented in Table 13.

Table 13: The classification results for different researches used the Coswara dataset to train and evaluate their models.

| Dataset                  | Research                                      | Performance (AUC)     |
|--------------------------|-----------------------------------------------|-----------------------|
| Coswara [9]              | University of Stellenbosch – (Pahar et al. 2020) [19] | 98% and 94%           |
|                          | Virufy – (Chaudhari et al. 2020) [18]          | 77.1%                 |
|                          | Wadhwani AI institute – (Bagad et al. 2020) [14] | 72%                   |
|                          | The proposed scheme                            | 96.4%                 |

7. Conclusion

In this study, we presented our methods for the sound-based COVID-19 diagnosis and detection using different respiratory sounds and human speech recorded by smartphones. For this purpose, we implemented deep learning-based models to distinguish COVID-19 from healthy audio samples. To train our models we used the Coswara dataset; An open dataset that contains 9 types of respiratory sounds for each user. We extracted 49 features from every audio sample to create 9 different datasets of features. The features we extracted were (MFCC, ZCR, Spectral Bandwidth, Spectral Centroid, Spectral Rolloff, RMS Energy), besides other statistical features that made a clear impact on the model performance. Like most COVID-19 datasets, Coswara is imbalanced with a smaller amount of positive samples, so we used Smote oversampling technique to balance the positive samples to the negatives in the training data. Then we used a simple deep learning model to train 9 classifiers on 9 datasets of features belongs to different sound types. Our results showed that the models trained and evaluated individually using Heavy Cough, Shallow Breathing, or Vowel O datasets, achieved a better performance than other models. But a combination of the models (Shallow Breathing, Heavy Cough, Shallow Cough, Fast Counting, Vowel E, Vowel O) could achieve an AUC of 96.4%. While using a combination of the models (Heavy Cough, Shallow Cough, Vowel E) achieved an AUC of 92%. Therefore, we emphasize the importance of using different respiratory sounds, coughs, and speech to train COVID-19 classifiers rather than using cough sounds only to build accurate and more trustworthy COVID-19 diagnostic and prescreening tools.

References

[1] C.-C. Lai, T.-P. Shih, W.-C. Ko, H.-J. Tang, P.-R. Hsueh, Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges, International Journal of Antimicrobial Agents. 55 (2020) 105924. https://doi.org/10.1016/j.ijantimicag.2020.105924.
[2] WHO Coronavirus (COVID-19) Dashboard, (2021). https://covid19.who.int/ (accessed May 1, 2021).

[3] J.B. Case, E.S. Winkler, J.M. Errico, M.S. Diamond, On the road to ending the COVID-19 pandemic: Are we there yet?, Virology. 557 (2021) 70–85. https://doi.org/10.1016/j.virol.2021.02.003.

[4] E.S. Goudouris, Laboratory diagnosis of COVID-19, J Pediatr (Rio J). 97 (2021) 7–12. https://doi.org/10.1016/j.jped.2020.08.001.

[5] D. Wang, B. Hu, C. Hu, F. Zhu, X. Liu, J. Zhang, B. Wang, H. Xiang, Z. Cheng, Y. Xiong, Y. Zhao, Y. Li, X. Wang, Z. Peng, Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China., JAMA. 323 (2020) 1061–1069. https://doi.org/10.1001/jama.2020.1585.

[6] A. Carfi, R. Bernabei, F. Landi, Persistent Symptoms in Patients After Acute COVID-19., JAMA. 324 (2020) 603–605. https://doi.org/10.1001/jama.2021.1426.

[7] G. Deshpande, B. Schuller, An Overview on Audio, Signal, Speech, & Language Processing for COVID-19, 2020.

[8] B. Schuller, D. Schuller, K. Qian, J. Liu, H. Zheng, X. Li, COVID-19 and Computer Audition: An Overview on What Speech & Sound Analysis Could Contribute in the SARS-CoV-2 Corona Crisis, Frontiers in Digital Health. 3 (2021). https://doi.org/10.3389/fdgth.2021.64906.

[9] N., P. Krishnan, R. Kumar, S. Ramoji, S.R. Chetupalli, N. R., P.K. Ghosh, S. Ganapathy, Coswara — A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis, Interspeech 2020. (2020). https://doi.org/10.21437/interspeech.2020-2768.

[10] L. Orlandic, T. Teijeiro, D. Atienza, The COUGHVID crowdsourcing dataset: A corpus for the study of large-scale cough analysis algorithms, 2020.

[11] A. Imran, I. Posokhova, H.N. Qureshi, U. Masood, M.S. Riaz, K. Ali, C.N. John, M.I. Hussain, M. Nabeel, AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app, Inform Med Unlocked. 20 (2020) 100378–100378. https://doi.org/10.1016/j.imu.2020.100378.

[12] C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthasombat, D. Spathis, T. Xia, P. Cicuta, C. Mascolo, Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data, Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. (2020). https://doi.org/10.1145/3394486.3412865.

[13] COVID-19 Sounds App, (2021). https://www.covid-19-sounds.org/en/ (accessed May 1, 2021).

[14] P. Bagad, A. Dalmia, J. Doshi, A. Nagrani, P. Bhamare, A. Mahale, S. Rane, N. Agarwal, R. Panicker, Cough Against COVID: Evidence of COVID-19 Signature in Cough Sounds, 2020.

[15] B. Subirana, F. Hueto, P. Rajasekaran, J. Laguarta, S. Puig, J. Malvehy, O. Mitja, A. Trilla, C.I. Moreno, J.F.M. Valle, A.E.M. González, B. Vizmanos, S. Sarma, Hi Sigma, do I have the Coronavirus?: Call for a New Artificial Intelligence Approach to Support Health Care Professionals Dealing With The COVID-19 Pandemic, 2020.

[16] J. Laguarta, F. Hueto, B. Subirana, COVID-19 Artificial Intelligence Diagnosis Using Only Cough Recordings, IEEE Open Journal of Engineering in Medicine and Biology. 1 (2020) 275–281. https://doi.org/10.1109/OJEMB.2020.3026928.

[17] A. Pal, M. Sankarasubbu, Pay Attention to the cough: Early Diagnosis of COVID-19 using Interpretable Symptoms Embeddings with Cough Sound Signal Processing, 2020.
[18] G. Chaudhari, X. Jiang, A. Fakhry, A. Han, J. Xiao, S. Shen, A. Khanzada, Virufy: Global Applicability of Crowdsourced and Clinical Datasets for AI Detection of COVID-19 from Cough, 2021.

[19] M. Pahar, M. Klopper, R. Warren, T. Niesler, COVID-19 Cough Classification using Machine Learning and Global Smartphone Recordings, 2020.

[20] H. Coppock, A. Gaskell, P. Tzirakis, A. Baird, L. Jones, B.W. Schuller, End-2-End COVID-19 Detection from Breath & Cough Audio, (2021).

[21] Fakhry, Ahmed, Jiang, Xinyi, Xiao, Jaclyn, Chaudhari, Gunvant, Han, Asriel, Khanzada, Amil, Virufy: A Multi-Branch Deep Learning Network for Automated Detection of COVID-19, (2021). https://doi.org/arXiv:2103.01806.

[22] B. McFee, C. Raffel, D. Liang, D. Ellis, M. Mevcicar, E. Battenberg, O. Nieto, librosa: Audio and Music Signal Analysis in Python, 2015. https://doi.org/10.25080/Majora-7b98e3ed-003.

[23] G. Peeters, A large set of audio features for sound description (similarity and classification) in the CUIDADO project, (2004).

[24] Spectral centroid for audio signals and auditory spectrograms, (2021). https://www.mathworks.com/help/audio/ref/spectralcentroid.html#mw_a6642d87-adb0-4199-bc69-60e08779270d (accessed May 1, 2021).

[25] A. Klapuri, M. Davy, Signal processing methods for music transcription, Springer Science & Business Media, 2007.

[26] Spectral spread for audio signals and auditory spectrograms, (2021). https://www.mathworks.com/help/audio/ref/spectralspread.html?s_tid=doc_ta#mw_35dea1a0-4ec6c-926f-5b7a5adcc593 (accessed May 1, 2021).

[27] Spectral rolloff point for audio signals and auditory spectrograms, (2021). https://www.mathworks.com/help/audio/ref/spectralrolloffpoint.html?searchHighlight=spectral%20rolloff&s_tid=srchtitle#mw_ad32153c-76d3-4c70-8c20-cf6de59c3b97 (accessed May 1, 2021).

[28] coefficient of variation, (2021). https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.variation.html (accessed May 1, 2021).

[29] S. Kokoska, D. Zwillinger, CRC standard probability and statistics tables and formulae, Crc Press, 2000.

[30] G. Lemaître, F. Nogueira, C.K. Aridas, Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning, Journal of Machine Learning Research. 18 (2017) 1–5.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research. 12 (2011) 2825–2830.

[32] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Y. Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, Xiaoqiang Zheng, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015. https://www.tensorflow.org/.
Conflict of Interest and Authorship Conformation Form

Please check the following as appropriate:

- All authors have participated in (a) conception and design, or analysis and interpretation of the data; (b) drafting the article or revising it critically for important intellectual content; and (c) approval of the final version.

- This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.

- The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript.

  - The following authors have affiliations with organizations with direct or indirect financial interest in the subject matter discussed in the manuscript:

| Author’s name | Affiliation |
|---------------|-------------|
|               |             |
|               |             |
|               |             |
|               |             |
|               |             |
|               |             |
|               |             |
|               |             |