Covid-19 Audio Classification

INTRODUCTION:

Cough audio signal classification has been successfully used to diagnose a variety of respiratory conditions, and there has been significant interest in leveraging Machine Learning (ML) to provide widespread COVID-19 screening. However, there is currently no validated database of cough sounds with which to train such ML models. The COUGHVID dataset provides over 20,000 crowdsourced cough recordings representing a wide range of subject ages, genders, geographic locations, and COVID-19 statuses. First, we filtered the dataset using our open-sourced cough detection algorithm. Second, experienced pulmonologists labeled more than 2,000 recordings to diagnose medical abnormalities present in the coughs, thereby contributing one of the largest expert-labeled cough datasets in existence that can be used for a plethora of cough audio classification tasks. Finally, we ensured that coughs labeled as symptomatic and COVID-19 originate from countries with high infection rates, and that their expert labels are consistent. As a result, the COUGHVID dataset contributes a wealth of cough recordings for training ML models to address the world's most urgent health crises.

Background & Summary :

One of the most common symptoms of COVID-19 is a dry cough, which is present in approximately 67.7% of cases5 . Cough sound classification is an emerging field of research that has successfully leveraged signal processing and artificial intelligence (AI) tools to rapidly and unobtrusively diagnose respiratory conditions like pertussis6 , pneumonia and asthma7 using nothing more than a smartphone and its built-in microphone. Several research groups have begun developing algorithms for diagnosing COVID-19 from cough sounds8, 9 . One such initiative, AI4COVID8 , provides a proof-of-concept algorithm but laments the lack of an extensive, labeled dataset that is needed to effectively train Deep Learning (DL) models. There are several existing COVID-19 cough sound datasets used to train Machine Learning (ML) models. Brown et al.9 have amassed a crowdsourced database of more than 10,000 cough samples from 7,000 unique users, 235 of which claim to have been diagnosed with COVID-19. However, the authors have not automated the data filtering procedure and consequently needed to endure the time-consuming process of manually verifying each recording. Furthermore, this dataset is not yet publicly available and therefore cannot be used by other teams wishing to train their ML and DL models. The Coswara project10, on the other hand, has publicly provided manual annotations of the crowdsourced COVID-19 coughs they have received, but as of September 2020, their dataset contains just slightly over 1,3000 samples. An alternative approach to crowdsourcing is the NoCoCoDa11, a database of cough sounds selected from media interviews of COVID-19 patients. However, this database only includes coughs from 10 unique subjects, which is not enough for AI algorithms to successfully generalize to the global population. In this work, we present the COUGHVID crowdsourcing dataset, which is an extensive, validated, and publicly-available dataset of cough recordings. With more than 20,000 recordings – 1,010 claiming to have COVID-19 – originating from around the world, it is the largest known COVID-19-related cough sound dataset in existence. In addition to publicly providing our cough corpus, we have trained and open-sourced a cough detection ML model to filter non-cough recordings from the database. Furthermore, we have undergone an additional layer of validation whereby 3 expert pulmonologists annotated a fraction of the arXiv:2009.11644v1 [cs.SD] 24 Sep 2020 dataset to

determine which crowdsourced samples realistically originate from COVID-19 patients. In addition to COVID-19 diagnoses, our expert labels and metadata provide a wealth of insights beyond those of existing public cough datasets. These datasets either do not provide labels or contain a small number of samples. For example, the Google Audio Set12 contains 871 cough sounds, but it does not specify the diagnoses or pathologies of the coughs. Conversely, the IIIT-CSSD13 labels coughs as wet vs dry and short-term vs long-term ailments, but it only includes 30 unique subjects. The COUGHVID dataset contributes over 2,000 expert-labeled coughs, all of which provide a diagnosis, severity level, and whether or not audible health anomalies are present, such as dyspnea, wheezing, and nasal congestion. Using these expert labels along with subject metadata, our dataset can be used to train models that detect a variety of subjects' information based on their cough sounds. Overall, our dataset contains samples from a wide array of subject ages, genders, COVID-19 statuses, pre-existing respiratory conditions, and geographic locations, which potentially enable AI algorithms to successfully perform generalization. Finally, we assert the validity of our data by ensuring that samples labeled as COVID-19 originate from countries where the virus was prevalent at the time of recording, and that the experts exhibit a reasonable degree of agreement on the cough diagnoses. The first step to building robust AI algorithms for the detection of COVID-19 from cough sounds is having a reliable, high-quality dataset, and the COUGHVID dataset effectively meets this pressing global need.

Problem Statement(summary):

As we have all came out from the major outbreak of covid pandemic, some of our researchers are still working on building the efficient models to predict the covid-19 and related pandemic diseases without using any sensory touch to the patients, by using the audio clips of the patients we can derive some features about the cough audio and these features can help us to predict whether the patient is covid positive or negative.

Data Collection: All of the recordings were collected between April 1st, 2020 and September 10th, 2020 through a Web application deployed on a private server located at the facilities of the École Polytechnique Fédérale de Lausanne (EPFL), Switzerland. The application was designed with a simple workflow and following the principle "one recording, one click", according to which if someone simply wants to send a cough recording, they should have to click on no more than one item. The main Web interface has just one "Record" button that starts audio recording from the microphone for up to 10 seconds. Once the audio recording is completed, a small questionnaire is shown to get some metadata about the age, gender, and current condition of the user, but even if the questionnaire is not filled, the audio is sent to the server. The variables captured in the questionnaire are described in Table 1. Also, the user is asked for permission to provide their geolocation information, which is not mandatory. Finally, since coughing is a potentially dangerous activity in the scope of a global pandemic, we provide easy-to-follow safe coughing instructions, such as coughing into the elbow and holding the phone at arm's length, that can be accessed from the main screen. Database Cleaning A common pitfall of crowdsourced data is that it frequently contains samples unrelated to the desired content of the database. In order to allow users of the COUGHVID database to quickly exclude non-cough sounds from their analyses, we developed a classifier to determine the degree of certainty to which a given recording constitutes a cough sound. These output probabilities of the classifier were subsequently included in the metadata of each record under the cough_detected entry. To train the classifier, we first hand-selected a set of 121 cough

sounds and 94 non-cough sounds including speaking, laughing, silence, and miscellaneous background noises. These recordings were preprocessed by lowpass filtering (fcuto f f = 6 kHz) and downsampling to 12 kHz. Next, 68 audio features commonly used for cough classification were extracted from each recording. 40 of the features are those described by Pramono et al.6 , and the implementation of our 19 Energy Envelope Peak Detection features is detailed by Chatrzarrin et al.14. We also included a signal length feature, as well as the power spectral density (PSD) of the signal in 8 hand-selected frequency bands. These bands were chosen by analyzing the PSDs of cough vs non-cough signals and selecting the frequency ranges with the highest variation between the two classes. Finally, we trained an eXtreme Gradient Boosting (XGB)15 classifier using 78% of the available data. The hyperparameters of the XGB model were tuned using Tree-structured Parzen Estimators (TPE)16 with a precision objective using the training data for 10-fold, 20%-test shuffle-split cross-validation. The ROC curve of the cough classifier is displayed in Figure 1, which users of the COUGHVID database can consult to set a cough detection threshold that suits their specifications. As this figure shows, only 10.4% of recordings with a cough_detected value less than 0.8 actually contain cough sounds. Therefore, they should be used only for robustness assessment, and not as valid cough examples.

Metadata

Dataset consists of the following files:

- **train.csv** - the training set
- **test.csv** - the test set
- **sample_submission.csv** - a sample submission file in the correct format
- **original.csv** - supplemental information about the original raw data
  use of external data is allowed.
  refer the discussion forum section of the competition to understand more about the dataset.

Sample of dataset:

4e47612c-6c09-4580-a9b6-2eb6bf2ab40c.json

{ "datetime": "2020-04-10T10:30:31.576207+00:00",

"cough_detected": "0.9466",

"age": "50",

"gender": "male",

"respiratory_condition": "True",

"fever_muscle_pain": "False",

"status": "COVID-19",

"expert_labels_1": { "quality": "ok", "cough_type": "dry", "dyspnea": "False", "wheezing": "False", "stridor": "False", "choking": "False", "congestion": "False", "nothing": "True", "diagnosis": "COVID-19", "severity": "mild" } }

Objective

- Create an audio classifier for Covid-19 classification
- Create an audio regressor for grading Covid-19 severity

Creating an untrained neural network to feed cough audio. The cough audio will be evaluated between values of 0 to 1 and a threshold would be set for determining the probability of covid-19 infection. An audio regressor would then help in grading the level of covid-19 infection.

Models Used:

1) Convolutional Neural Network (CNN): Convolutional neural networks (CNN) are one of the most popular models used today. This neural network computational model uses a variation of multilayer perceptrons and contains one or more convolutional layers that can be either entirely connected or pooled. These convolutional layers create feature maps that record a region of image which is ultimately broken into rectangles and sent out for nonlinear processing.

Advantages:

- Very High accuracy in image recognition problems.
- Automatically detects the important features without any human supervision.
- Weight sharing.

Disadvantages:

- CNN do not encode the position and orientation of object.
- Lack of ability to be spatially invariant to the input data.
- Lots of training data is required.

2) Artificial Neural Network (ANN):

Artificial Neural Network (ANN), is a group of multiple perceptrons or neurons at each layer. ANN is also known as a Feed-Forward Neural network because inputs are processed only in the forward direction.
This type of neural networks are one of the simplest variants of neural networks. They pass information in one direction, through various input nodes, until it makes it to the output node. The network may or may not have hidden node layers, making their functioning more interpretable.

• Input layer: The first layer is input layer which contains mfcc extracted features that is passed through model. The layer has input node of 1000, with an input shape of (40,) because during feature extraction, n mfcc was 40. Activation function is ‘ReLu’.

• Five hidden layers: The nodes in five hidden layers decrease in order of 750, 500, 250, 100 & 50. ‘ReLu’ activation functions are used which will help in tackling the vanishing gradient problem. The negative part of the argument will be removed by using the ’ReLu’ activation function. o f ReLU(x)=max(0, x)

• Output layer: The final layer is the output layer. The activation function used is ’SoftMax’. The output layer will be used to generate the output based on the number of labels, i.e., 10. SoftMax will give a probability of each class that will sum up equal to 1. The optimizer used was ’Adam’. Adam optimizing algorithm is an extension of stochastic gradient descent, which is used to update the weights of the network in the training data.

Advantages:

- Storing information on the entire network.
- Ability to work with incomplete knowledge.
- Having fault tolerance.
- Having a distributed memory.

Disadvantages:

- Hardware dependence.
- Unexplained behavior of the network.
- Determination of proper network structure.

3)**KNN:**

Advantages:-

1. **No Training Period**- KNN modeling does not include training period as the data itself is a model which will be the reference for future prediction and because of this it is very time efficient in term of improvising for a random modeling on the available data.

2. **Easy Implementation**- KNN is very easy to implement as the only thing to be calculated is the distance between different points on the basis of data of different features and this distance can easily be

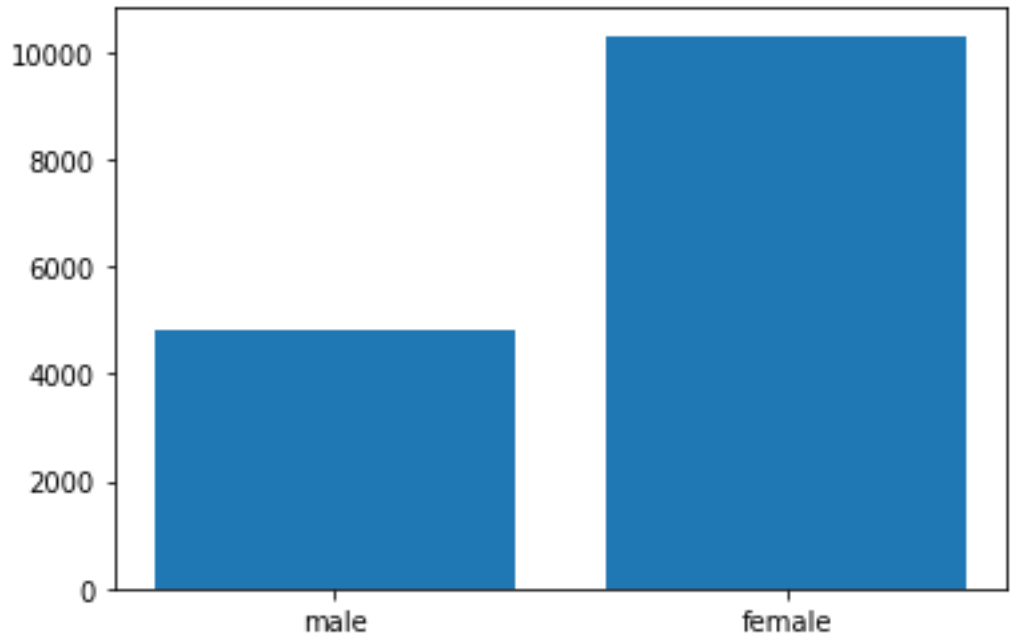
calculated using distance formula such as- Euclidian or Manhattan

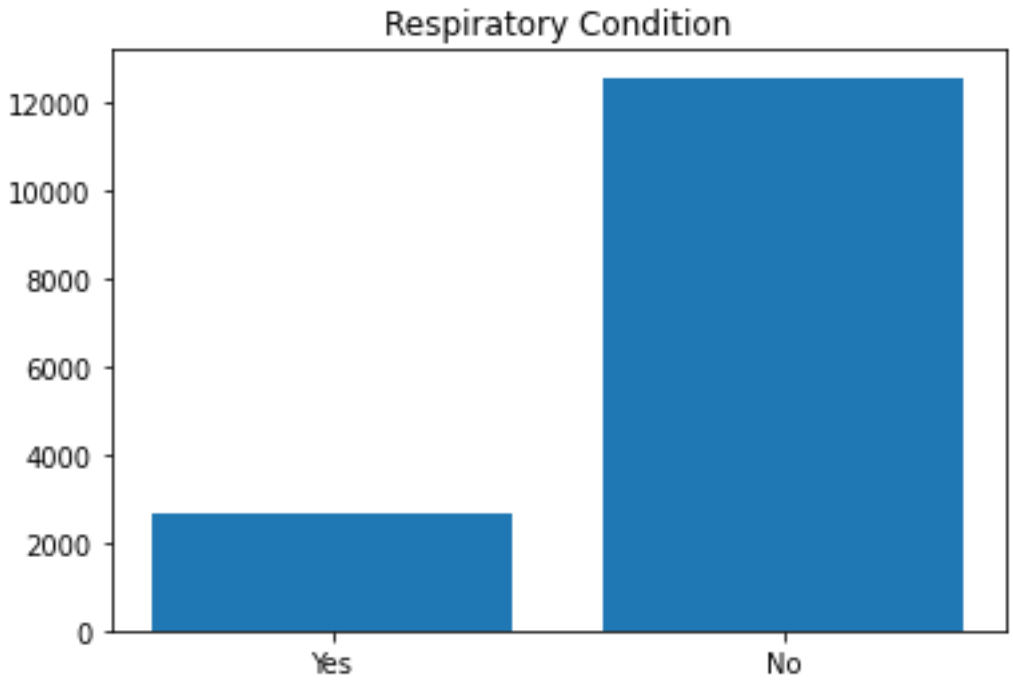
3. As there is no training period thus new data can be added at any time since it wont affect the model.

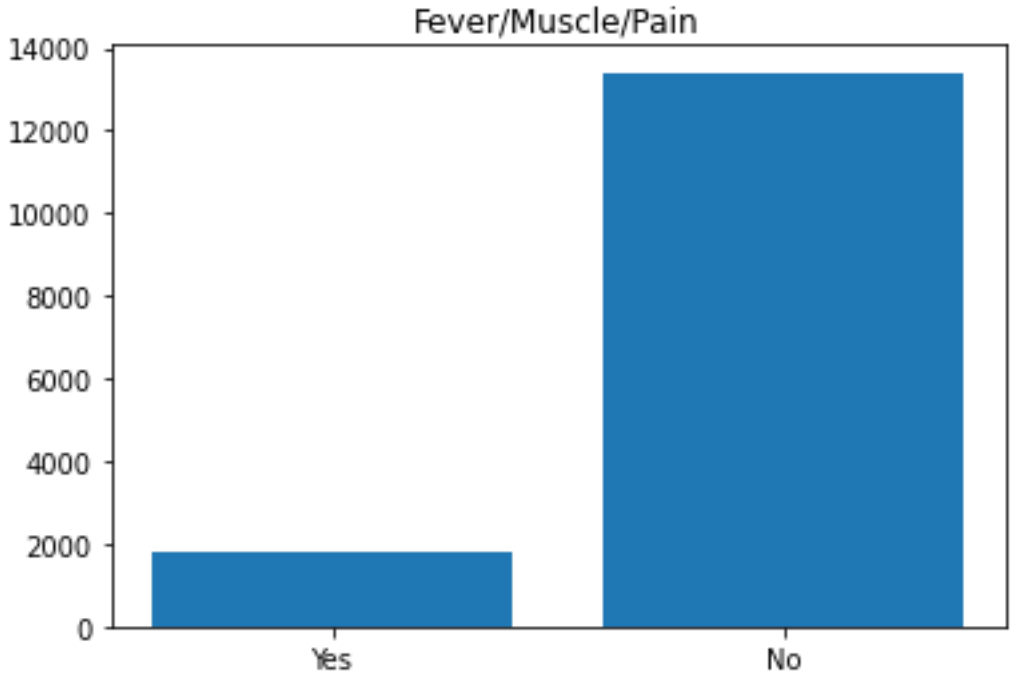
Disadvantages:-

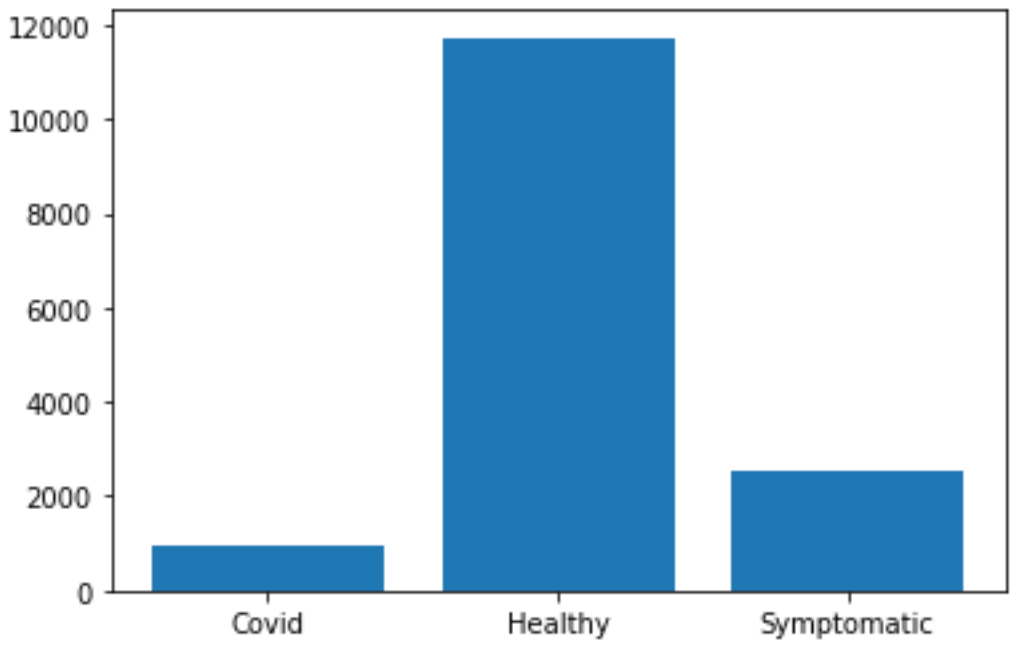
1. **Does not work well with large dataset** as calculating distances between each data instance would be very costly.
2. **Does not work well with high dimensionality** as this will complicate the distance calculating process to calculate distance for each dimension.
3. **Sensitive to noisy and missing data**
4. **Feature Scaling-** Data in all the dimension should be scaled (normalized and standardized) properly .

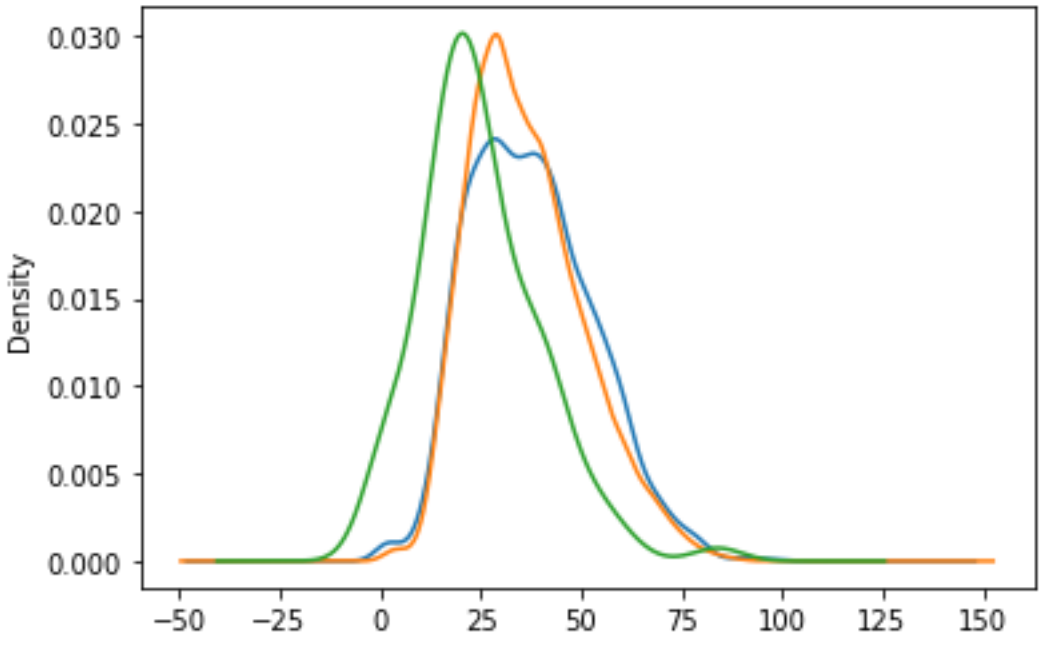
Model Information:

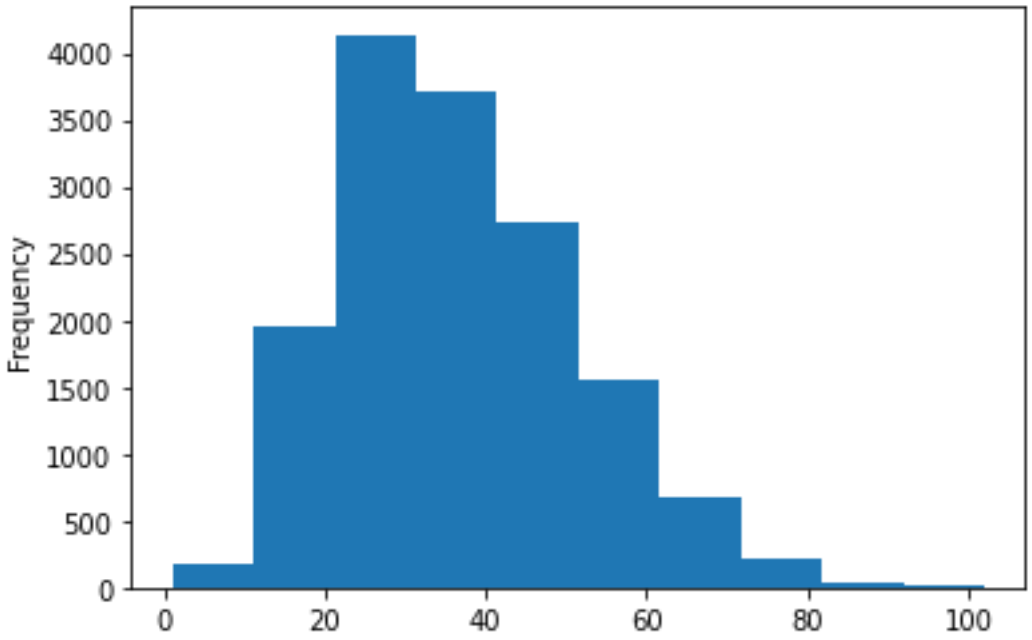


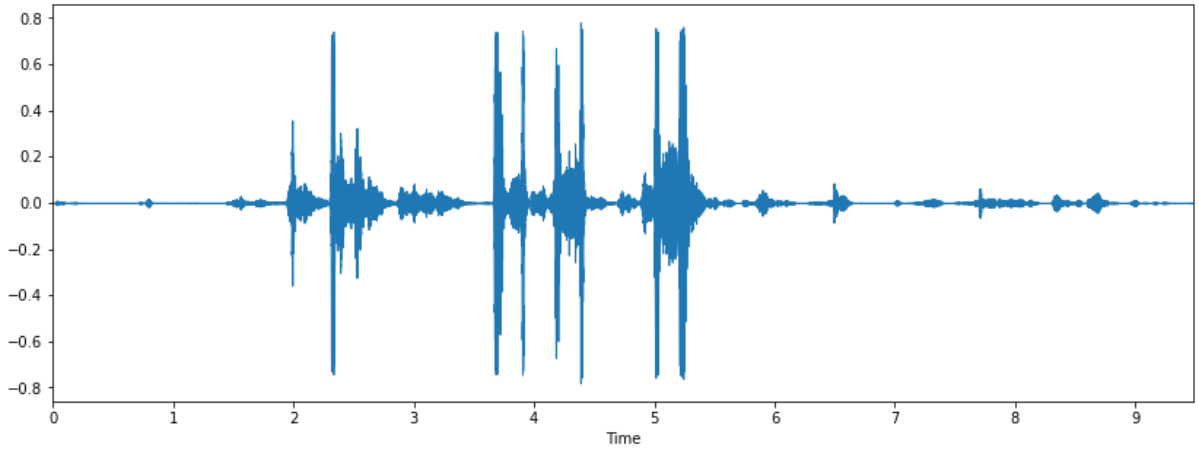

Respiratory Condition
12000
10000
8000
6000
4000
2000
0
Yes
No

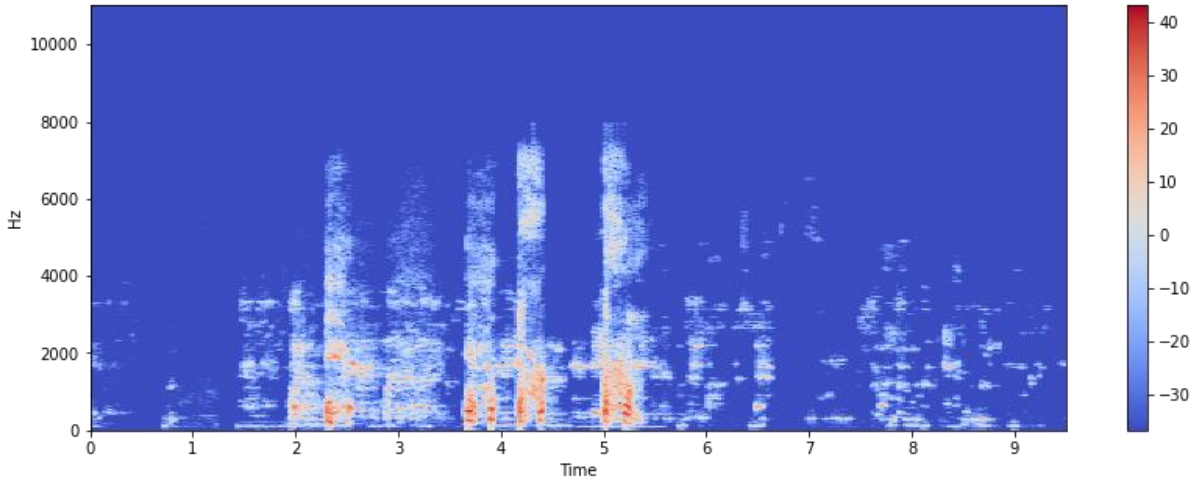

Fever/Muscle/Pain
14000
12000
10000
8000
6000
4000
2000
0
Yes
No

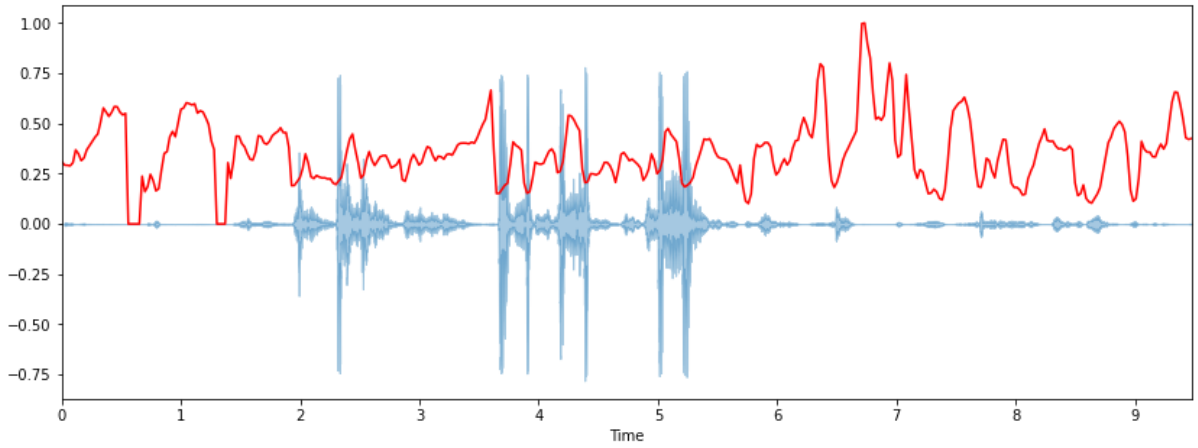

12000
10000
8000
6000
4000
2000
0
Covid
Healthy
Symptomatic


0.030
0.025
0.020
0.015
0.010
0.005
0.000
Density
−50
−25
0
25
50
75
100
125
150


Frequency
4000
3500
3000
2500
2000
1500
1000
500
0
0
20
40
60
80
100


0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
0
1
2
3
4
5
6
7
8
9
Time


Hz
10000
8000
6000
4000
2000
0
0
1
2
3
4
5
6
7
8
9
Time
40
30
20
10
0
−10
−20
−30



Model prediction example:

```
Covid-19   Healthy   Symptomatic
1/1 [==============================] - 0s 51ms/step


[[0.06420393 0.76825404 0.16754209]]
```

Accuracy:

```
TEST ACCURACY: 0.7808803915977478
```