University of Essex
Department of Computer Science andElectronic Engineering

MSc DISSERTATION

AI-based trusted decision for detection of COVID-19 using audio recording

Prepared by
Dhanalakshmi Kurunguntla
Student ID: 2111477

Supervisor: **Dr Ali Zulfiqar**

December 14, 2022

Colchester

Contents

List of Figures

ABSTRACT:

Coronavirus disease 2019 (COVID-19) is a viral infection brought on by the coronavirus 2, a virus that causes severe acute respiratory syndrome (SARS-CoV-2). A new coronavirus strain that hasn't been seen in humans before. Viral Testing(PCR ,Antigen) and antibody Testing are methods available to detect the COVID-19. My primary goal of dissertation to find other methods to detect COVID-19 in a patient to assist in the diagnosis of those who are infected and stop the spread. The aim of the project is to implement machine learning model for detection of COVID-19 using the audio files. The classification of cough audio signals has been used successfully to diagnose the different types of respiratory conditions. To enable comprehensive COVID-19 screening, there has been a lot of interest in using machine learning (ML). created machine learning models to detect the COVID-19 used the binary classification. More than 20,000 cough records representing age, gender, geographic location, and COVID-19 status are available from dataset COUGHVID . The AI model trained on cough detection achieved 81% the accuracy of random forest and Ensemble model got the best accuracy (validation) 89.34% compared with SVM which got validation accuracy of 89.46 %,logistic regression which the validation accuracy of 89.65%.

1. Introduction:

A zoonotic disease is an infectious disease brought on by the spread of pathogens from animals to humans. Coronavirus is a zoonotic virus. Corona is a type of infection that causes respiratory problems in people[2]. In Wuhan, China, in December 2019, an unidentified animal transmitted the coronavirus (COVID-19), which was first found in bats and transmitted to humans[3]. On January 30, 2020, the World Health Organization recognized COVID-19 as a Public Health Emergency of International Concern. On March 11, 2020, the organization declared it a global pandemic. Figure1 shows the current cases in covid19 worldwide.

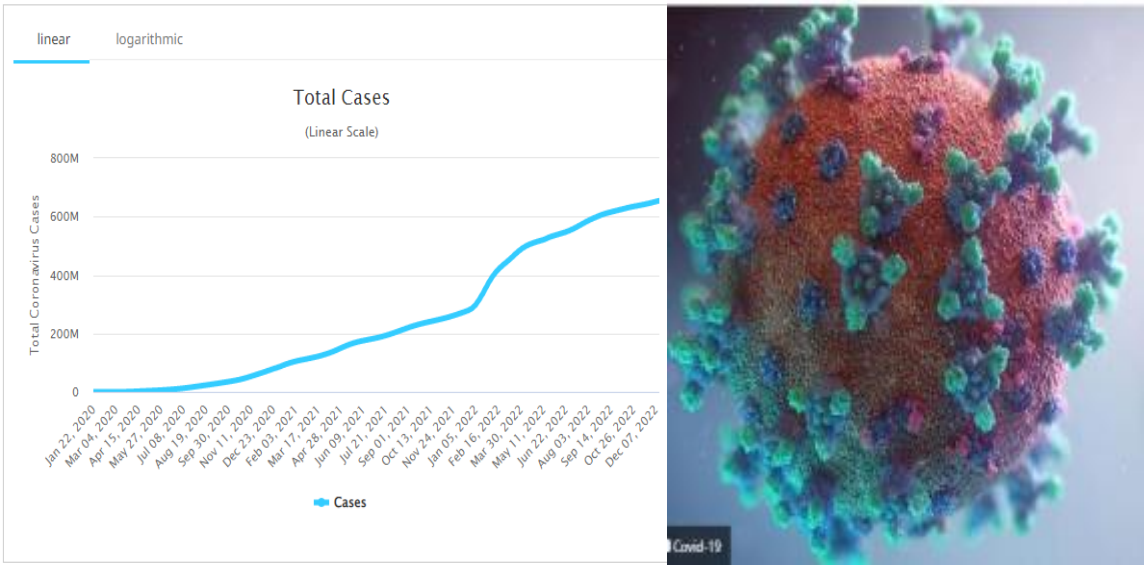

Figure1 Cases in total (worldwide) 13/12/2022

This virus may affect those who experience high temperatures, coughing, sore throats, fatigue, muscle aches, throat pain, and difficulty breathing.Figure2 shows the symptoms of covid19.

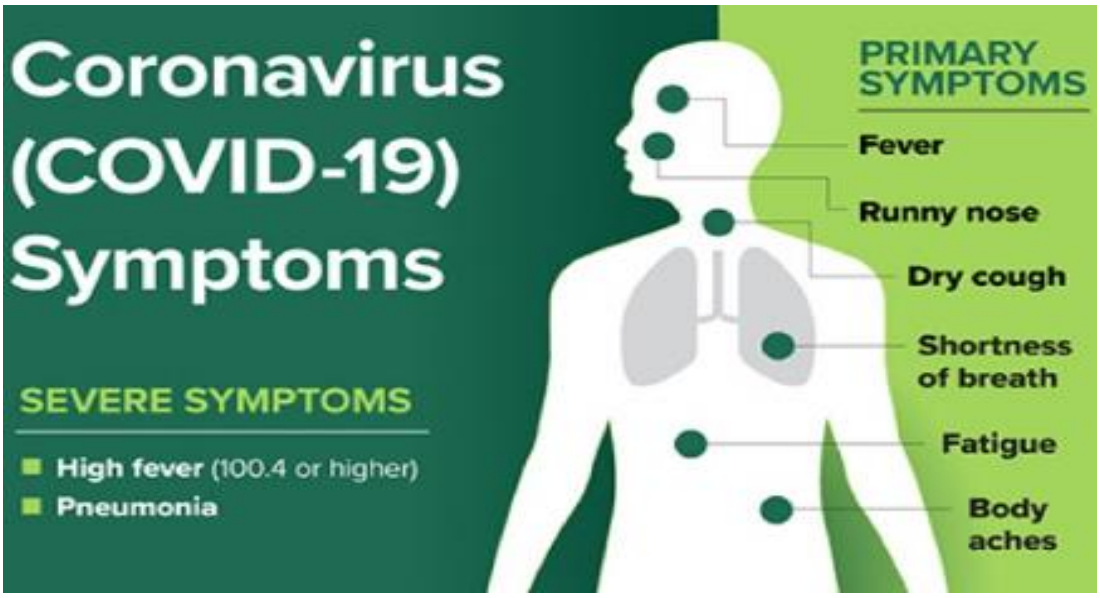

Figure 2: symptoms of covid19

A social, economic, and health crisis has been brought on by the COVID-19 issue. Lockdowns have caused many to lose their jobs. These issues have put a great deal of strain on people's mental health and negatively impacted their physically well-being. The ability to operate remotely thanks to the internet has helped the sector survive. However, the sectors that have been severely impacted entail physical labor, such as technicians and labour. The lockdowns have a particularly negative impact on people who work in low-wage such as street labourers [1]. In the last ten years, the fields of AI and its applications have experienced tremendous growth. The level of intelligence that robots possess is rising, and they are now more capable of making wise decisions than before. This implies that robots are losing their innocence and that, as a result of this development, they run a higher chance of being used maliciously. For instance, there has been a lot of discussion on the ethical implications of AI in the medical field. How much confidence would you place in a machine to determine a patient's diagnosis for cancer based on their symptoms? How would a machine account for false positives and/or false negatives that match a syndrome? How would a machine distinguish between asthma and a reactive airway disease, both of which call for different treatments? These are some of the frequently asked questions. This is because most AI algorithms depend on data, and applications that use wrong or insufficient data may not succeed. Positively, AI has demonstrated potential in simple clinical testing and opened the door for more intricate study on statistical techniques used in healthcare.

Additionally, a lot of research is being done in areas of medical diagnosis where it has been demonstrated that it is challenging and challenging for humans to complete. For instance, researchers are currently successfully using breast cancer detection [17] and they assert to have a 99% success rate in identifying the tumour's malignancy or benignity. The goal of this thesis is to significantly advance the field of health technology that could be used to help find novel coronaviruses using cough sounds.

Reverse Transcription Polymerase Chain Reaction (RT-PCR) testing is not widely available to the worldwide population, requires direct human interaction to administer, and has a variable turnaround time. One way to slow the exponential increase in COVID-19 cases is to create a model that can conduct biological testing without requiring a less number of human contact t. As a result, numerous AI-based programs that rely more on audio and minimal human interaction have been utilized for testing and the early diagnosis of respiratory disorders [4].

2.Focus of the project and research questions:

2.1 Challenges:

COVID-19 has been spreading around the world since the beginning of 2020, resulting in approximately 650 million cases and 6.5 million deaths. we are using vaccination to protect ourselves from corona virus. Most of the recent research focuses to find other methods to detect COVID-19 in a patient to assist in the diagnosis of those who are infected s, whether it is positive or negative and stop the spread. The breath or cough of people can be recorded by several audio applications that have been released. The "Coughvid" [], "Breath for Science" [] "Coswara" [], and "CoughAgainstCovid" [] are a few examples. With the availability of these datasets, numerous studies have been conducted that use machine learning and cough and/or breath signals to detect the virus. It is essential that we can create models that can categorise the healthier and covid people from the available datasets data.

2.2 Project objectives:

The project concentrated to implement a classifier using machine learning algorithms that can accurately classify to detect the covid-19 by research and analysis of previously employed methods. Models are chosen and utilized to classify COVID-19 and healthy based on the literature review.

2.3 Question for research:

The results from this project will be able to address the issues raised by the question below: How effective are machine learning algorithms are identifying the decision for detection of COVID?

To determine the covid19 status, several machine learning models are created to train on the COUGHVID dataset, which offers over 25,000 crowdsourced cough recordings reflecting a wide range of participant ages, genders, geographic locations, and statuses. Every model created will be evaluated based on how well it categorizes the covid state and how well it generalizes on test data. The next of the paper is structured as follows: Chapter 3 summarizes the literature review, Chapter 4 Methodology used in this project, Chapter 5 discussed about the results obtained, Chapter 6 explain the differences between the various models and Chapter 7 Future work.

3.LITERATURE REVIEW:

The second wave of the COVID-19 pandemic has directly contributed to the rising number of persons who have passed away from their illnesses. As is evident, the second wave is causing havoc on the healthcare systems of numerous countries. In order to stop the virus from spreading, regional frequent testing and contact tracing may take the place of regional restrictions.

The "Trace, Test, and Treat" approach has also flattened the pandemic trajectory in its early stages (for example, in Singapore, South Korea, and China). The infection rate must therefore be brought under control in order to reduce the burden on medical resources, and quick and affordable COVID-19 infection detection methods are an urgent necessity. The affected countries have implemented a number of precautionary steps to impede the disease's spread. These tactics include encouraging people to keep their distance from one another and practise good personal hygiene, enhancing infection screening systems through the application of multi-functional testing, pursuing mass vaccination to lower the possibility of a pandemic occurring in the future, and many other related techniques. Developing or underdeveloped nations are still trying to improve their COVID-19 detection capacities because the current methods, like reverse transcription-polymerase chain reaction (RT-PCR), are expensive kits that must be used for on-site testing and because these kits are not always simple to obtain.

Therefore, for identifying and diagnosing COVID-19 and stopping local epidemics of COVID-19 infection, low-cost pre-screening assays that are simple to distribute, dependable, and accurate are essential. In addition to the conventional diagnostic strategy of RT-PCR, a number of AI-based techniques have recently been developed. To distinguish COVID-19 from other bacterial and viral infections, these techniques make use of chest X-rays and CT scans. However, one must travel to a testing facility or well-equipped clinical facilities in order to use RT-PCR, CT scans, and X-rays for the purpose of diagnosis. Given the contagious nature of COVID-19 and the number of people working together in close quarters during the testing procedure, there is a considerable risk that the infection will spread more widely. Making a model that can do biological testing without involving a lot of people could be one solution to the issue of how to stop the exponential growth of COVID-19 cases. In order to test for and make an early diagnosis of respiratory illnesses, a variety of AI-based apps have been used. These applications employ audio and require less human interaction than conventional techniques.

Using AI-based diagnostic models, a patient's cough symptoms can be used to determine the presence of asthma, pulmonary emphysema, tuberculosis, pneumonia, and whooping cough. Cough is a recognisable sign of numerous respiratory illnesses. The signs of coughing have been used to identify many respiratory illnesses. It is commonly known that COVID-19 infection may affect a person's respiratory system, which may affect their coughing, breathing, and voice tone. The presence of COVID-19 infection has been detected using AI models that are based on audio in a number of recent studies. peech and voice related digital biomarkers have recently become a topic of study interest in the domains of medicine and artificial intelligence due to the ease of signal gathering and accurate diagnostic outcomes (AI).In this section we will discuss about the few of the researchers done the research for detecting the covid19 using the Artificial intelligence.

This section summarizes the classification study done by earlier researchers to identify the COVID19.

The authors of this paper (Madhurananda Pahar, Marisa Klopper, Robin Warren, Thomas Niesler) [5] used the Coswara dataset (92 COVID-19 positive individuals and 1079 healthy subjects) and second dataset were gathered in South Africa (contains 18 COVID-19 positive and 26 COVID-19 negative). MFCCs characteristics were retrieved from the cough audio using a feature extraction method that keeps the time-domain patterns. Authors[5] trained and tested the Seven machine learning classifiers logistic regression (LR), k-nearest neighbour (KNN), support vector machine (SVM), multilayer perceptron (MLP), convolutional neural network (CNN), long short-term memory (LSTM), using the cross-validation method and a residual-based neural network architecture (Resnet50). The results demonstrate that while all classifiers could recognise COVID-19 coughs, the Resnet50 classifier performed the best, with an area under the ROC curve (AUC) of 0.98 being the best at differentiating between COVID-19 positive and healthy coughs. This kind of cough audio classification is affordable and simple to use, making it potentially a valuable and practical method of non-contact COVID-19 screening.

The authors of [6] used convolutional neural networks (CNN) to create an automatic cough classifier. This classifier can identify whether coughing is present in a given audio clip. According to the study, their system produced an accuracy of 86.94%. This study also develops a CNN-based classifier to identify bronchiolitis, pertussis, and bronchitis, demonstrating the capability of AI-based models to identify respiratory illnesses from cough sounds. However, I have discovered two problems in this paper: The reproducibility of the results may be in doubt due to two factors: first, both models were trained using a small dataset; second, despite being

less complex, the cough detector is a classifier that needs a fixed short audio sample of five seconds to assess whether or not it has a cough. Long audio files must therefore go through preprocessing where they are compressed in order for the detector to work successfully. Additionally, the diagnostic model will be noisy if the background noise in a 5-second sample is longer than the cough itself, necessitating additional feature engineering to eliminate noise from the cough-containing audio clip.

Researchers at Cambridge University who are the authors of [7] are conducting research related to this one, and they have found preliminary encouraging results of diagnosing COVID-19 from speech and coughs. They used crowd-sourced data, extracting customized features in addition to features through transfer learning, and obtain an encouraging AUC of 80% for covid patients (despite the short sample size). However, as the authors note, more data are still being gathered for this study, which is still in its early phases. Furthermore, the results are not trustworthy enough to be used as a screening tool at this time because they are based on crowd-sourced data with self-reported classifications.

4.Methodology:

4.1 Dataset:

The purpose is to create a cough-based COVID-19 classifier, and the dataset includes both COVID-19-positive and non-COVID-19-positive cough samples. COUGHVID[8] is one of the largest COVID-19 cough datasets that is currently available , which is open source. Approximately 1010 samples of COVID-19 coughs are included in each of the roughly 20,000 recordings that make up this collection. From April 1st 2020 until September 10th 2020, these cough recordings were made. A Web application was created and placed on a private server inside the university with the aim of making the recordings easier to handle. Metadata is gathered for the following kinds of information:

1.What is the current state of your health? Three additional options are available for selection: Despite being in good health, my symptoms have not been accurately identified; Someone informed me that I have COVID-19.

2. Which of the following characteristics do you own? Additional respiratory illnesses include Do you suffer from a fever or muscle pain?

3. Gender,

4. Age.

As it is necessary to capture coughs in a safe environment, instructions are given on how to do so, including how to cough into your elbow and keep your phone at arm's length from your mouth. They created a classification system to determine the amount of coughing sounds included in an audio recording. The metadata JSON file can be located under the item for "cough detected," and it contains information about the likelihood of hearing a cough during each recording. Three very skilled pulmonologists were asked to manually annotate 1,000 cough recordings in order to get a better understanding of the datasets' quality.

The objective of this experiment was to learn the pulmonologists' perceptions of the cough recordings' quality and the self-labeling that was done during the recording. The selection of the 1000 cough recordings was done by stratified random sampling. The stratification was based on the self-reported labels. Twenty-five percent of the recordings contained COVID, thirty-five percent had symptoms, twenty-five percent were deemed to be in good health, fifteen percent had no reported status, and fifteen percent of these were examined by all three pulmonologists. The data had an equal representation of men and women, 65.5 to 33.8, respectively. 7.5% of the samples were considered to have COVID, 15.5% were seen to be symptomatic, and 77% were assessed to be healthy. This Coughs can be recorded by the programme CoughVid for ten seconds at a time. Users are permitted to cough as often as they like during this time without experiencing

any negative effects. Following an analysis of the dataset's metadata, it was found that 11% of the cough recordings were of poor quality, and that only 2% of the collection contained any cough recordings at all dataset has been used to train both our COVID-19 classifier and our cough detector. In order to make the annotations, it was decided to use the recordings in their original form. If there wasn't a cough, the sample would be recorded as lacking annotations. This was done in order to make the cough detector more durable. The process of data preparation continued with data annotation after doing an examination of the cough quality of the CoughVid datasets[8] .A cough detector was created with the goal of being able to establish the timestamps of coughing episodes that took place inside a sample. This serves as a filter to significantly lower the amount of background noise in a sample, which enhances the performance of any cough-based model, in this case the COVID-19 classifier. As a result, as the next step in the data preparation process, the dataset needs to be annotated in order to record timestamps of cough occurrences.

4.2 Audio/signal pre-processing:

The use of speech signals for COVID-19 identification can be a valuable and affordable technique because it does not require any difficult medical tests. With the use of this method's automatic detection tool, a patient's initial state can be quickly diagnosed without the need to visit a hospital or ask for any medical assistance.

Let's look at various audio features related to machine learning and audio processing.

4.2.1 spectrum and cepstrum:

Cepstrum and spectrum are two main components in audio processing.

The Fourier transform of a signal is a spectrum. A time-domain signal is converted to a frequency domain signal using a Fourier transform. In other words, a spectrum is a representation of the time-domain audio signal in the frequency domain. The inverse Fourier transform is used to create a cepstrum by first taking the spectrum's logarithmic magnitude. This produces a signal that is neither in the frequency domain (since we used an inverse Fourier transform) nor in the time domain (because we took the log magnitude prior to the inverse Fourier transform). The resulting signal's domain is referred to as the quefrency.[3]

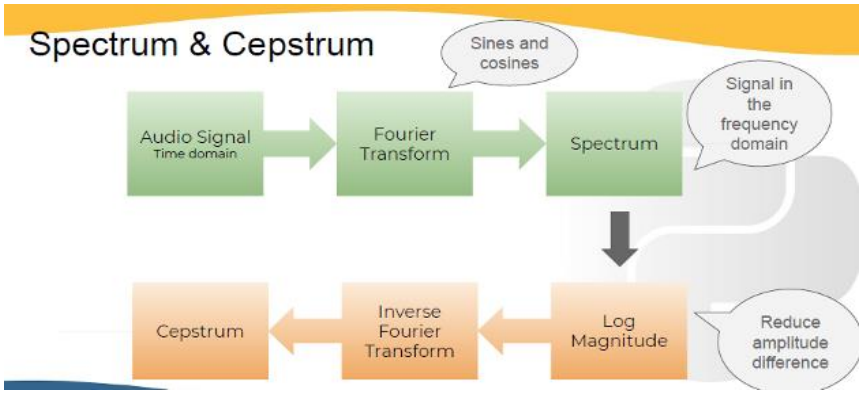

Figure 3 : spectrum and cepstrum

The frequency domain signal is related to ear biology. Before humans can analyse and comprehend a sound, several processes must take place. This occurs in the cochlea, a fluid-filled area of the ear that has many microscopic hairs connecting to nerves. Some of the hairs are rather short, while others are longer. The shorter hairs have a greater resonant frequency, while the longer hairs have a lower resonant frequency. Consequently, the ear functions as a natural Fourier transform analyzer.

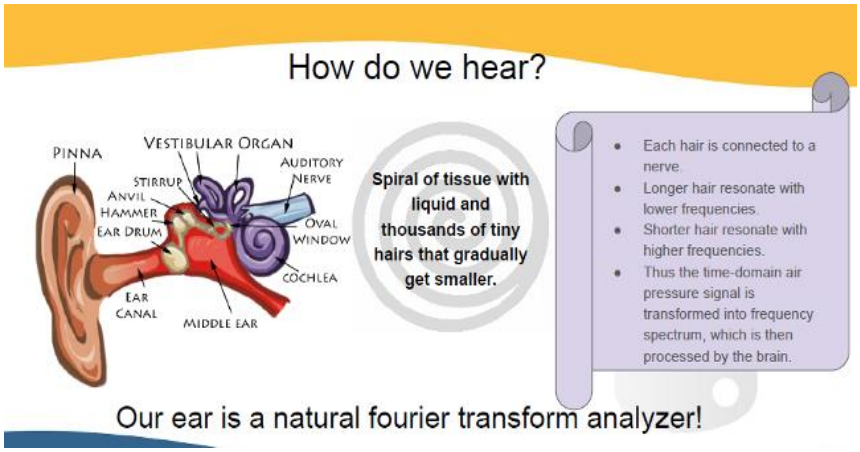

Figure 4 : Our ear is a natural fourier transform analyzer

4.2.2 Speech Extraction Features:

Coughing sounds, whose bandwidth and properties are significantly distinct from the entire speech signal, are usually used in the context of COVID-19 detection[9].

With the help of the inner ear, the deepest part of the ear, which is responsible for converting vibrations into electrical impulses that are then transmitted to the brain, humans have the innate ability to recognize sounds in their environment. Using a biomimicry approach, researchers were able to develop Feature Extraction Techniques, a crucial node in the audio classification pipeline

that plays a significant role as the inner ear in our bodies. Mel-frequency cepstral coefficients (MFCCs) and gamma-frequency cepstral coefficients are two methods for extracting features [10]

4.2.2.1 MFCC (Mel Frequency Cepstral Co-efficient):

The MFCC Mel Frequency Cepstral Co-efficient -based features are widely applied and very effective for speech assessment metrics, music information retrieval, and other applications .. Cepstrum provides information on the rate of change in spectral bands. Any periodic element (such as echoes) appears as sharp peaks in the associated frequency spectrum (i.e., Fourier spectrum) when time signals are conventionally analyzed. To get this, the temporal signal is subjected to a Fourier transform[11].

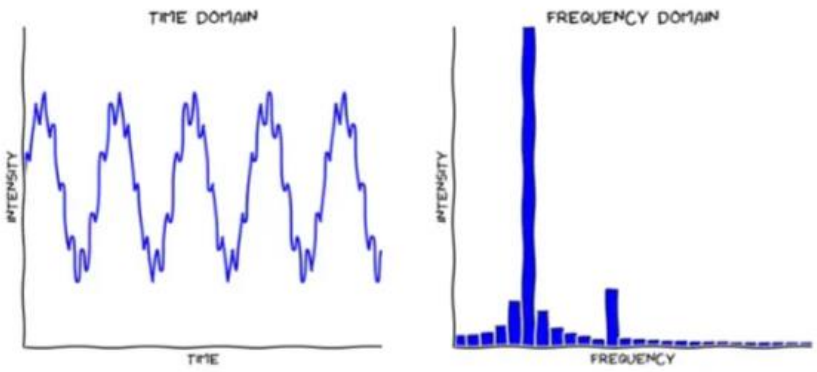

Figure 5 Time domain vs frequency domain

By first calculating the log of this Fourier spectrum's magnitude and then calculating the spectrum of this log through a cosine transformation. Wherever the original time signal contains a periodic component, we see a peak. The frequency spectrum we get as a result is neither in the frequency domain nor the time domain since we apply a transform on the frequency spectrum itself. cepstrum is the name of this spectrum, which is the logarithmic spectrum of the time signal.[10]

When cepstrum was initially developed, it was used to describe the seismic echoes produced by earthquakes. One of a speech signal's qualities, pitch is quantified as the frequency of the signal. Mel scale is a scale that connects a tone's perceived frequency to its actual measured frequency. It scales the frequency to match the range of human hearing more precisely.

Human hearing can detect frequencies between 20Hz and 20kHz. Think about a 300 Hz melody. This sounds quite like the dialler tone on a landline phone. Now imagine a 400 Hz melody (a little higher pitched dialer tone). Now, compare the distance between these two, however your brain

may detect it. Imagine a signal at 900 Hz (which sounds like the feedback from a microphone) and a sound at 1 kHz. Although the real difference between these two noises is the same, the perceived distance between them may seem to be more than it is (100Hz). Such variations are tried to be captured by the Mel scale[10][11].

The formula below can be used to translate a frequency measured in Hertz (f) to the Mel scale.

$$\mathrm{Mel}(f) = 2595 \log\left(1 + \frac{f}{700}\right)$$

Figure Melscale formula

The form of a human's vocal tract affects every sound they produce (including tongue, teeth, etc). Any sound that is made can be precisely represented if its shape can be identified. The

vocal tract is described by the envelope of the time power spectrum of the speech signal, and MFCC (which is nothing more than the coefficients that make up the Mel-frequency cepstrum) appropriately depicts this envelope.

A step-by-step description of MFCCs is shown in the block diagram below.

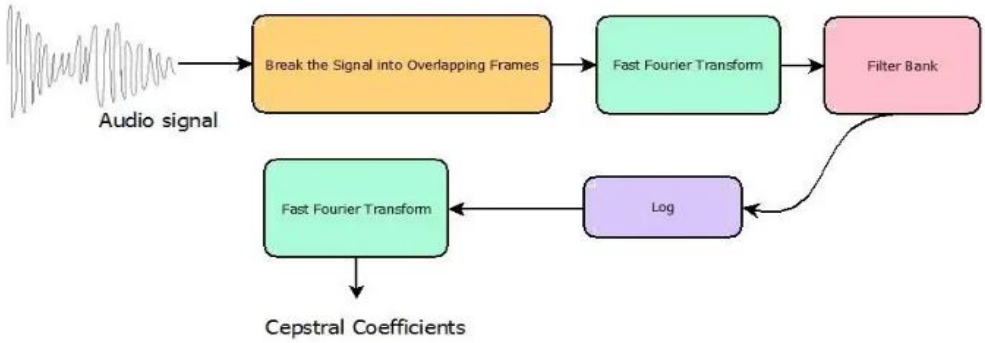

Figure 6 MFCC Process

Another observation regarding human hearing is that our ears start to become less sensitive to frequencies when sound frequency rises above 1 kHz. This is compatible with something known as the Mel filter bank.

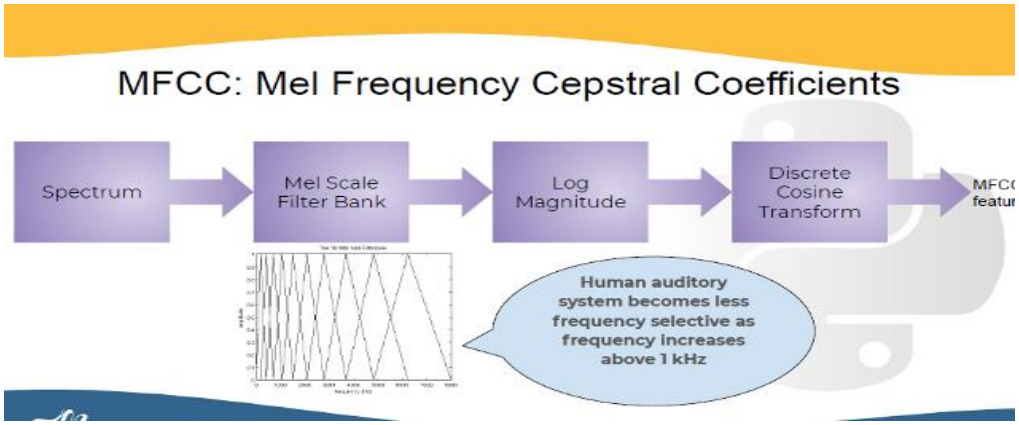

Figure 7 MFCC: Mel Frequency Cepstral Coefficients

The Mel cepstrum is generated by running a spectrum through the Mel filter bank, taking the log magnitude, and performing a discrete cosine transform (DCT). The primary information and peaks of the signal are extracted via DCT. JPEG and MPEG compressions both rely mainly on it. The audio information can be summarised by looking at the peaks. MFCCs are a general term for the first 13 coefficients extracted from the Mel cepstrum.[10][9][11]. These would be frequently used to train machine learning models.

Below are the steps that are usually used to compute the MFCC are as follows:
Apply windowing and frame pre-processing to the input signal, Determine the energy of the frame, Using the FFT technique, determine the discrete Fourier transform, by mapping the power spectrum onto the Mel scale and employing triangular overlapping windows, the Mel filter bank is applied, calculate the logarithm and finally Energy and DCT features can be combined to produce MFCC features.[12]

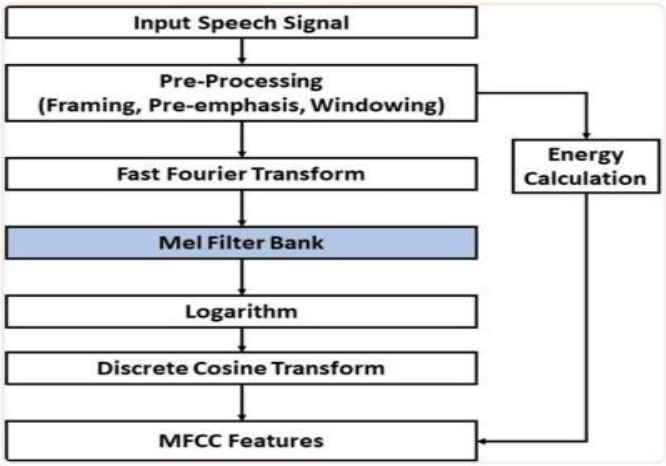

Figure 8 MFCC feature extraction

4.2.2.2 Gammatone-frequency cepstral coefficients (GFCCs):

Gammatone frequency cepstral coefficients are coefficients produced using a set of Gammatone filter banks. To produce those coefficients, we need a frequency-time representation of the signal known as a Cochlegram (relative to the Cochlea, a component of the inner ear), which can be obtained out of the Gammatone fliterbank.[13]

The calculation of the gammatone cepstral coefficients is similar to the MFCC extraction method. First, the audio signal is windowed into brief frames, typically lasting 10 to 50 ms.[6]. After that, the fast Fourier transform (FFT) of the signal is applied to the GT filter bank, which emphasizes the perceptually significant sound signal frequencies. This filter bank is made up of the frequency responses of several GT filters. Finally, the log function and discrete cosine transform (DCT) are used to represent the human loudness perception and de-correlate the outputs of the logarithmic compression filter, resulting in improved energy compaction.[13]

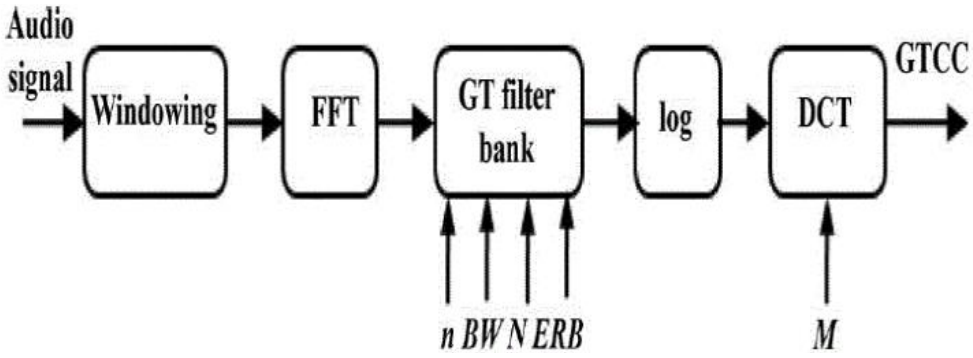

Figure 9 Block diagram of GTCC

Windowing: The audio samples are first windowed into 30 ms long frames with a 15 ms overlap (using a Hamming window).

The Hamming window equation is given as:
If the window is defined as W (n), $0 \leq n \leq N-1$ where
N = number of samples in each frame
Y[n] = Output signal
X (n) = input signal
W (n) = Hamming window, then the result of windowing signal is shown below:
$Y(n) = X(n) * W(n)$

Figure 10 Hamming window equation

GT Filter Bank: The frequency responses of numerous GT filters compose up the bank of GT filters. It is used to emphasize the perceptually significant sound signal frequencies in the fast Fourier transform (FFT) of the signal.

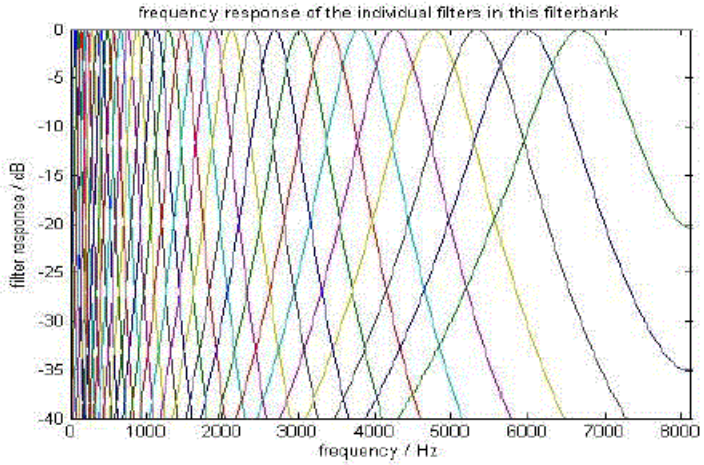

Figure 11 Filter bank output

Fast Fourier Transform: The transformation of each frame of N samples from time domain to frequency domain.

Y (w) = FFT [h (t) * X (t)] = H (w) * X(w) [13]
if X (w), H (w), and Y (w) represent, in that order, the Fourier Transforms of X (t), H (t), and Y (t) [13]

Discrete Cosine Transform (DCT):Better energy compaction is achieved by applying the discrete cosine transform (DCT) and the log function to replicate human loudness perception and de-correlate the outputs of the logarithmic-compressed filter.

4.3 Building a model:

Based on a survey of the literature, this models are chosen. By using the training data each model was trained and in training phase test data are used to estimate the performance of each model. performance metrics used are F1 score and accuracy and confusion matrix.

1. SVM
2. Logistic Regression
3. Ensemble Model
4. RandomForest

4.3.1 Support Vector Machine:

Use the supervised machine learning method known as the Support Vector Machine to address classification and regression problems. However, classification problems are where it is most useful. The SVM approach states that each data point is represented by a coordinate in an n-dimensional space, with each feature's value being the coordinate's value. Following that,

classification is performed by locating the hyper-plane that efficiently divides the two groups[14]. The location of a single observation is simply represented by a support vector. Between the two groups, the support vector machine classifier (SVM) is cutting edge (hyperplane and line). The support vector machine approach looks for a hyperplane in N-dimensional space (N is the number of features) that accurately classifies the data points. You can choose from a variety of hyperplanes to divide the data into two categories. The plane with the highest margin, which is the widest separation between data points in both classes, is where we want to look. By increasing the margin distance, more support is given, which makes it simpler to accurately classify following data points. We are able to maximise the classifier's margin by employing these training support vectors. Erase the support vectors to change the hyperplane's placement. These elements combine to form our SVM.[14]

4.3.2Logistic Regression:

The linear algorithm applied in this scenario is logistic regression, which is particularly effective at predicting issues involving binary class classification. Based on a set of provided features, it utilises a sigmoid function to forecast the likelihood of the class. Since the prediction is made as a probability, the output will fall between zero and one, and the direction of association can also be determined from LR.[15]

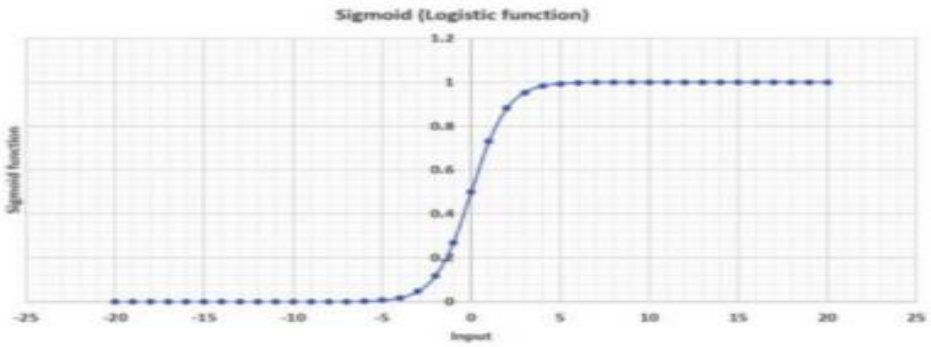

Figure 12 sigmoid (Logistic regession)

Use the supervised machine learning method known as the Support Vector Machine to address classification and regression problems. However, classification problems are where it is most useful. The SVM approach states that each data point is represented by a coordinate in an n-dimensional space, with each feature's value being the coordinate's value.

4.3.2 RandomForest:

The Bootstrap aggregating or bagging approach is used in randomforest to ensemble the numerous decision tree models. The forecasts are made using a variety of DT models in this method, and the final prediction is either reached by taking the mean of each individual prediction or by a majority vote.[16]

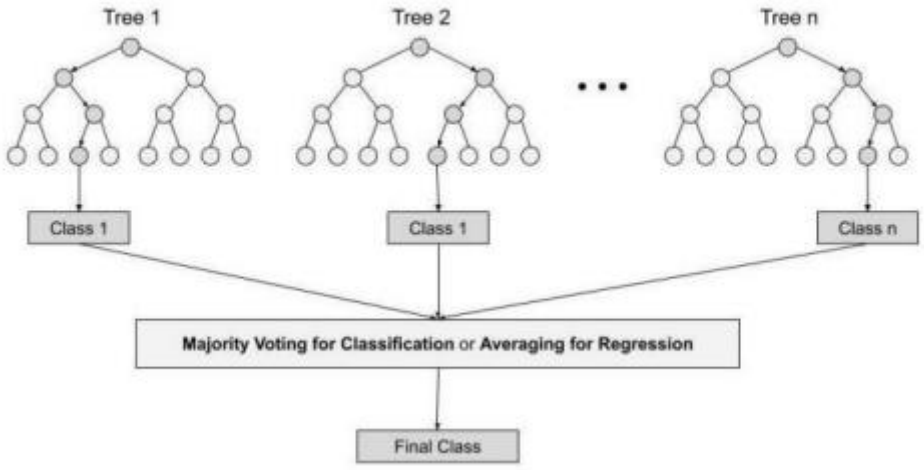

Figure 13 Classifier with Random Forests

Using Python program, I first identified how many uuids(unique Id given to each audio recording in the COUGHVID dataset) have a label in the given data. For this, first I have loaded metadata_compiled.csv into a dataframe. Then I created a function called 'multi_lebel_ids', which accepts this data frame and stores all the uuids with multilabels into one variable called 'multi_lebel_uuids' and remaining uuids where there is a single label are stored in to a variable called 'labeled_data'. 2841 audio files have labels and 2506 audios donot have any labels and 130 audio recordings which have muti_labels. Then I created two functions; one for extracting multilabels and other for extracting single labels. Both functions accepts the above created data frame and returns the expert labels, user labels, quality and severity in four different variables.

After this, I have created a function for extracting features from the audio recordings.
This function iterates over the dataframe and for each uuis in the dataframe, the function takes the audiofile from the path where all the audio recording are saved.The function uses MFCC library to extract the features from given audio recording file and check then checks whether the audio has mutilabels or mono label and accordingly sends the audio to either multi_label_selection_function or single_label_selection function. This process is repeated for all the audio recordings and from each audio files features are extracted and other important features such as age, gender, cough_detected,respiratory_condition,
fever_muscle_pain, status_SSL are taken for each audio recording and saved into a data frame called Tnsfd_data.

The data is then separated, all the features except the response/target variable are stored into a variable called 'X' and the target variable is stored in a variable called 'Y'.
The target variable in this case status and it holds two values; healthy and Covid-19. So, I created a function and transformed all the COVID-19 labels as 1 and Healthy as 0 and reassigned them to variable 'Y'. For handling missing values I have used SimpleImputer and all the NaN values are replaced by mean, implemented MinMax scaler for scaling the data.
Then I have split the data for training and testing. 75% of the data is used for training and 25% is used for testing. During training, I have used cross validation with 5 splits.

5.Results:

In comparison to earlier versions, the results of this study are generally modest. Although we have experimented with many techniques and parameters to determine how it would impact the performance of the model

The |below results were conducted using Matlab. The classification is done for two classes only “healthy” and “covid”. The following attributes are present in the dataset these are as follows:

- uuid
- datetime
- cough_detected
- latitude
- longitude
- age
- gender
- respiratory_condition
- fever_muscle_pain
- status

The predictors used are as follows:

- cough_detected
- age
- gender
- respiratory_condition
- fever_muscle_pain

The response variable used is :

- status

Model1 represents SVM, Model2 represents Logistic Regression and Model3 represents Ensemble Classifier.

Training Results :

Below are the results got after training model on training data in MATLAB.

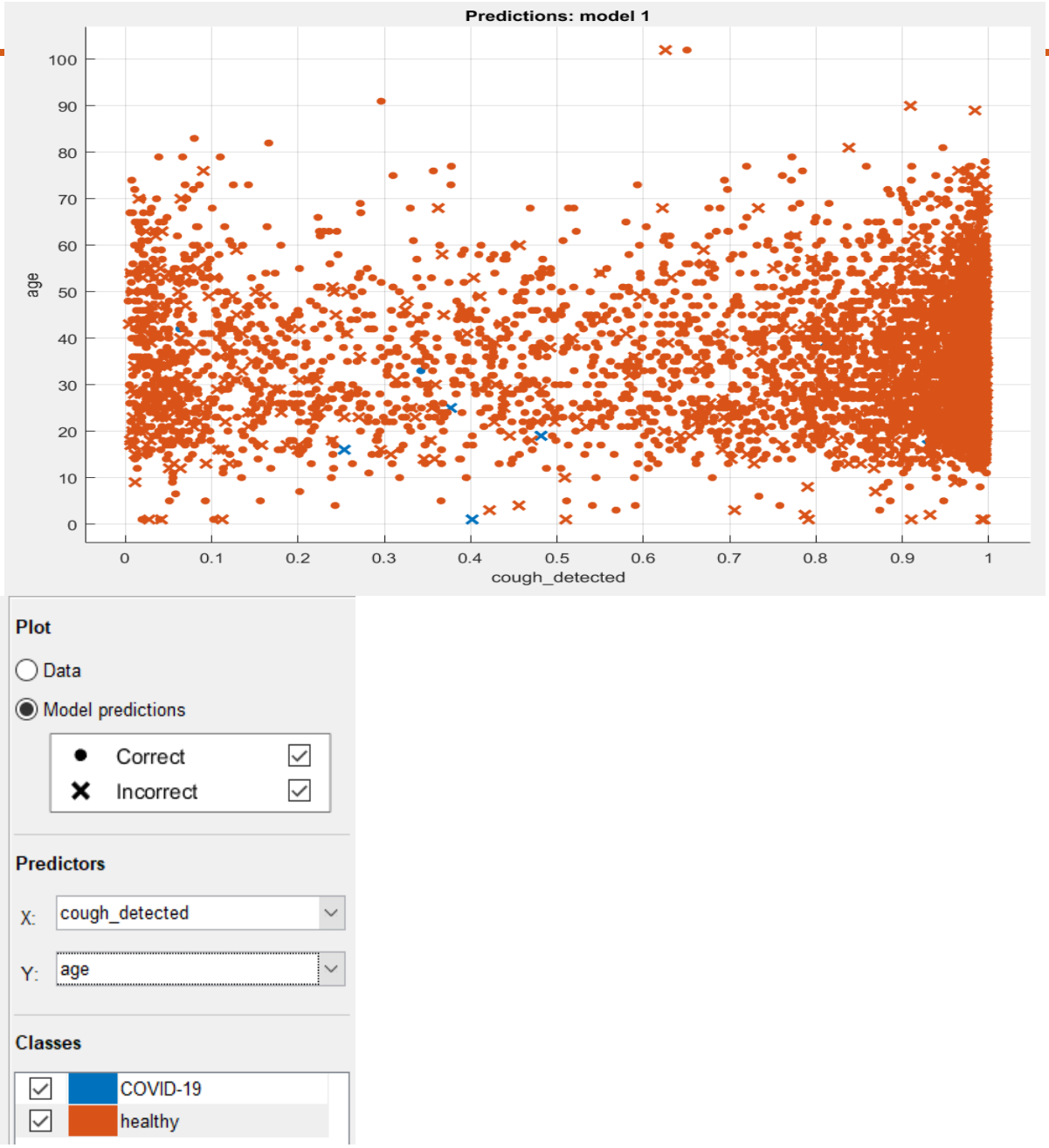

Figure 14 SVM Model predictions

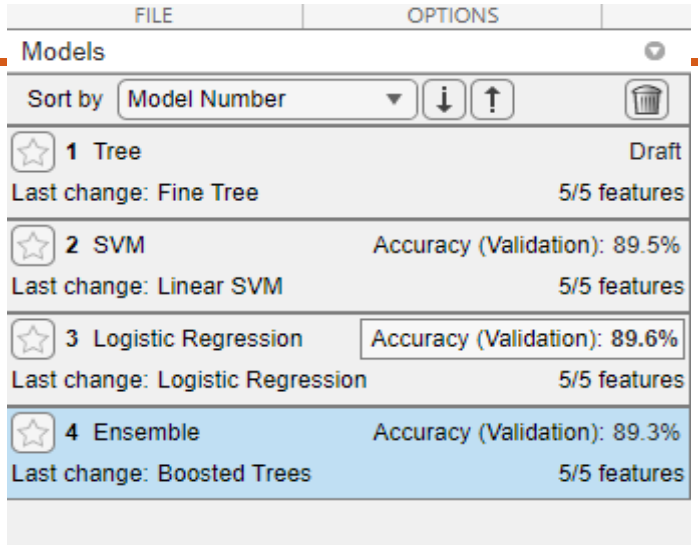

Fig 15 Testing results of SVM vs logistic regression vs Ensemble Algorithms

Below are the results got after training the models on training data in MATLAB. For SVM got the Accuracy (validation) 89.6%, logistic regression got the Accuracy ((validation) 89.6%,Ensemble got the accuracy ((validation)of 89.3%. Among all Ensemble model got the better accuracy than SVM and linear regression.

SVM:
Validated the confusion matrix ,ROC curve below are the results obtained in SVM

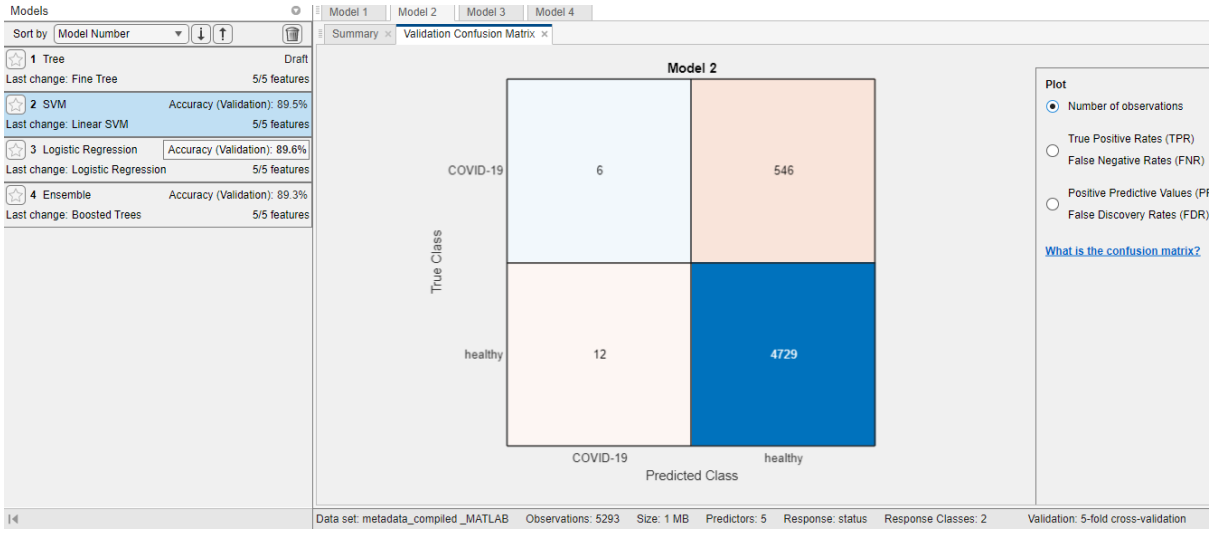

Figure16 SVM validation of confusion matrix No.of observations result

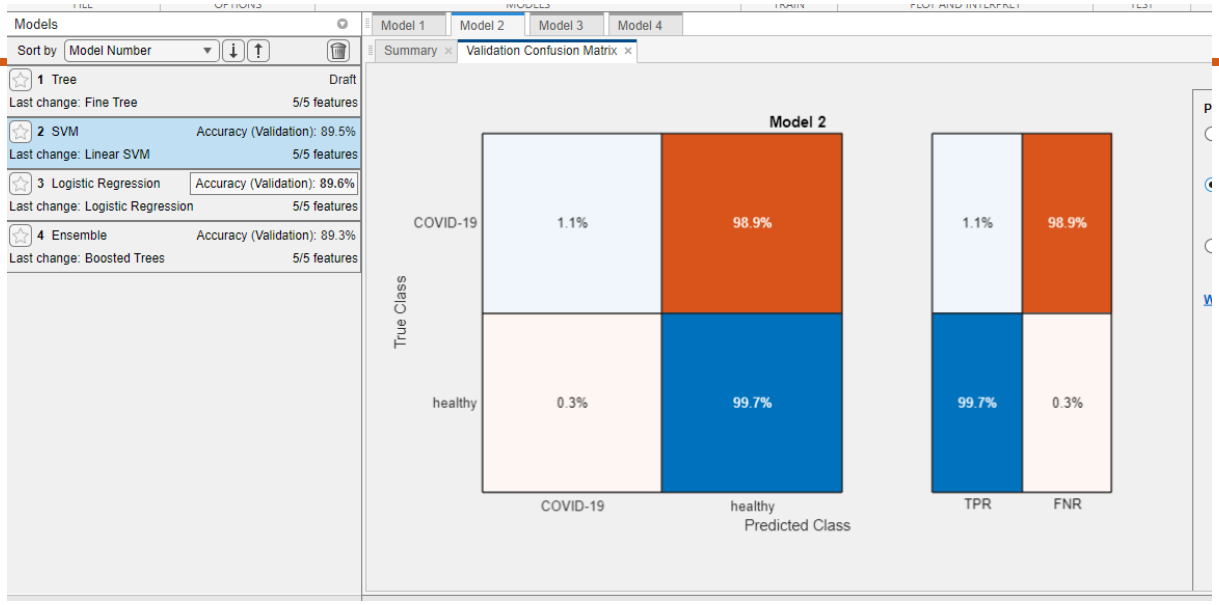

Figure17 SVM validation of confusion matrix result

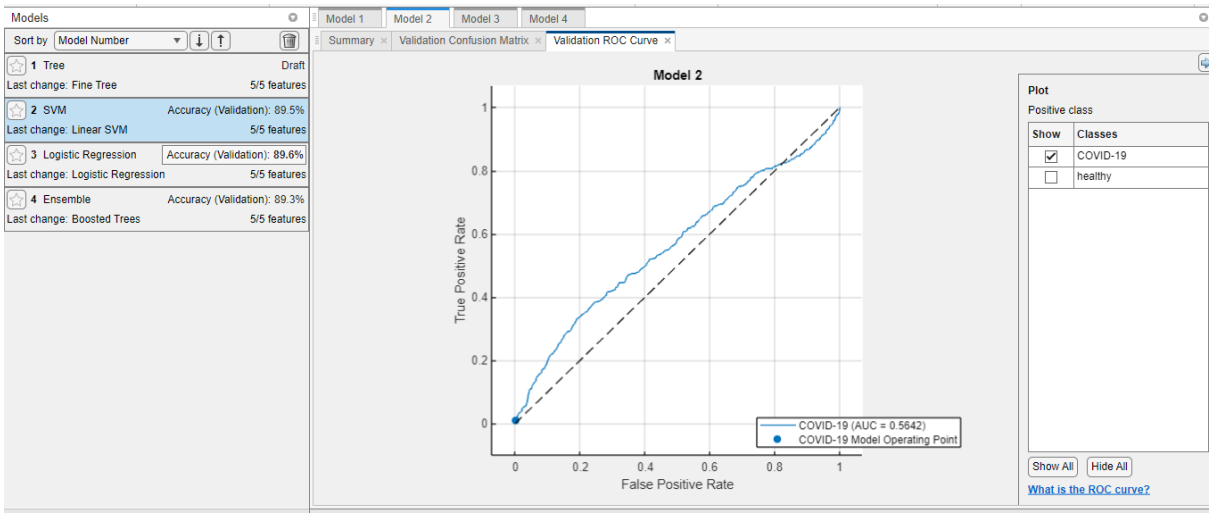

Figure18 SVM ROC curve result of COVID19 class

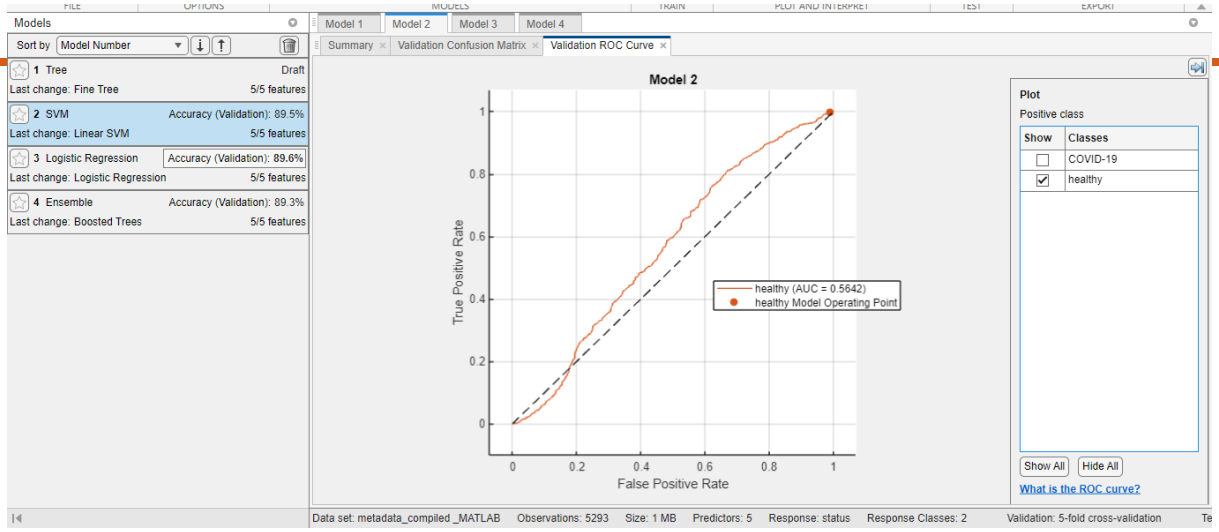

Figure19 SVM ROC curve result of healthy class

Linear regression:
Validated the confusion matrix ,ROC curve below are the results obtained in Linear regression after training the model .

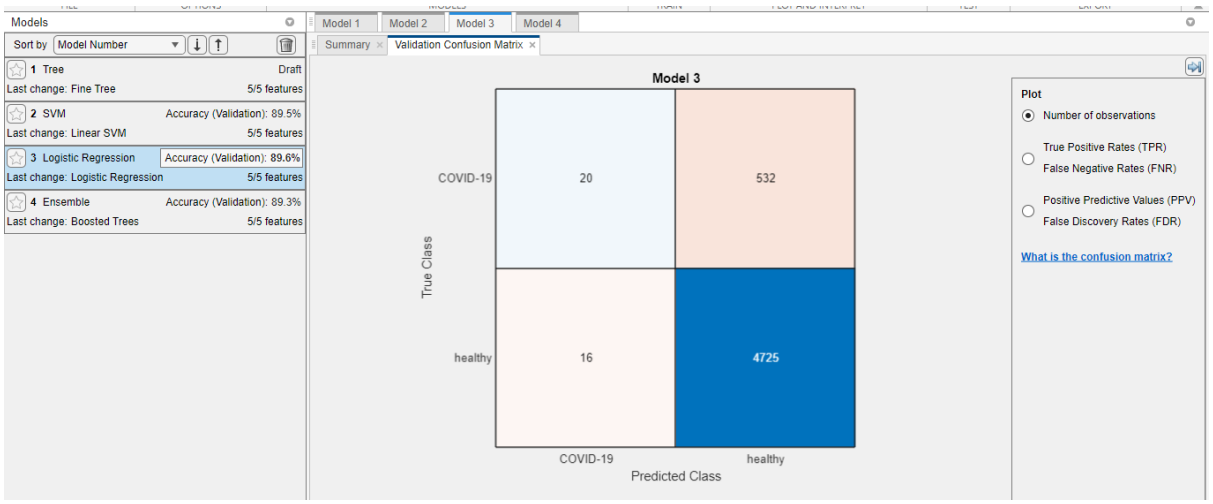

Figure 20 logistic regression confusion matrix no.of observation result

Figure 21 logistic regression confusion matrix result

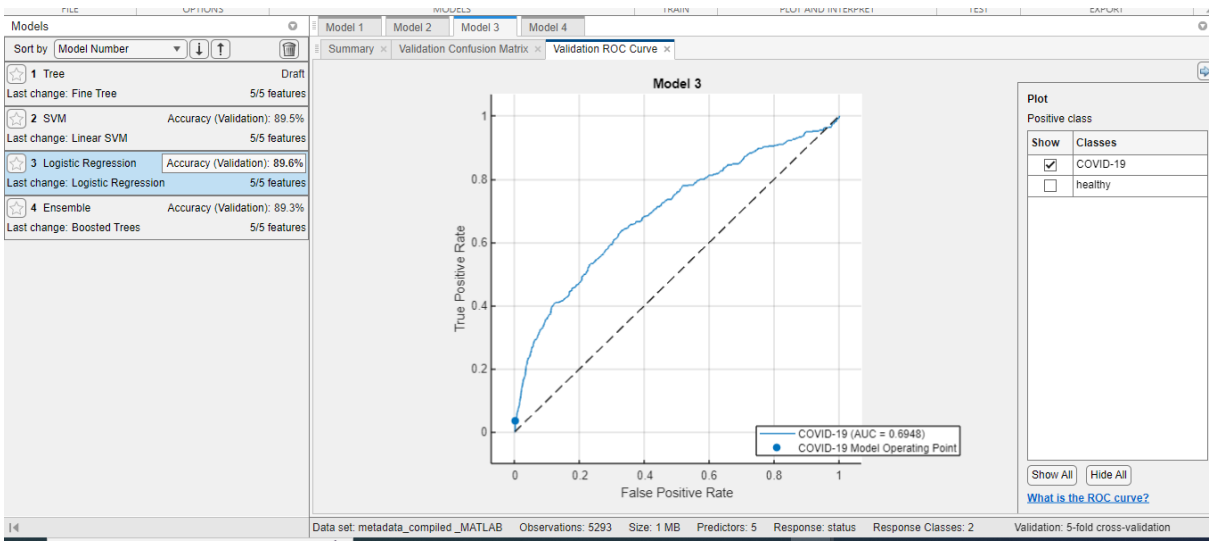

Figure22 logistic regression ROC curve result of Covid-19 class

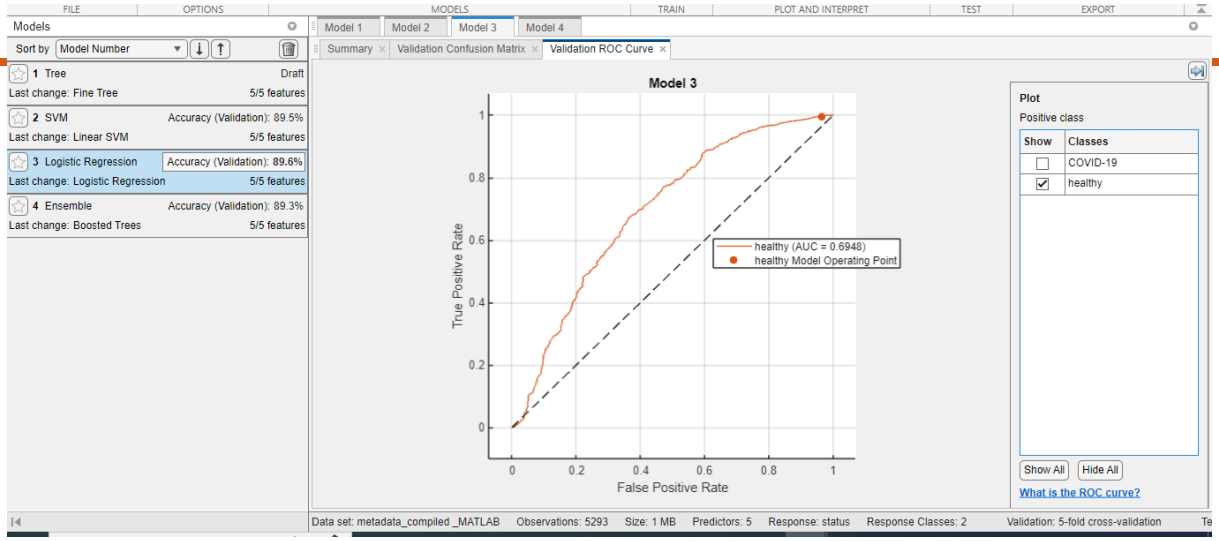

Figure 23 logistic regression ROC curve result of healthy class

Ensemble:

Validated the confusion matrix ,ROC curve below are the results obtained in Ensemble after training the model .

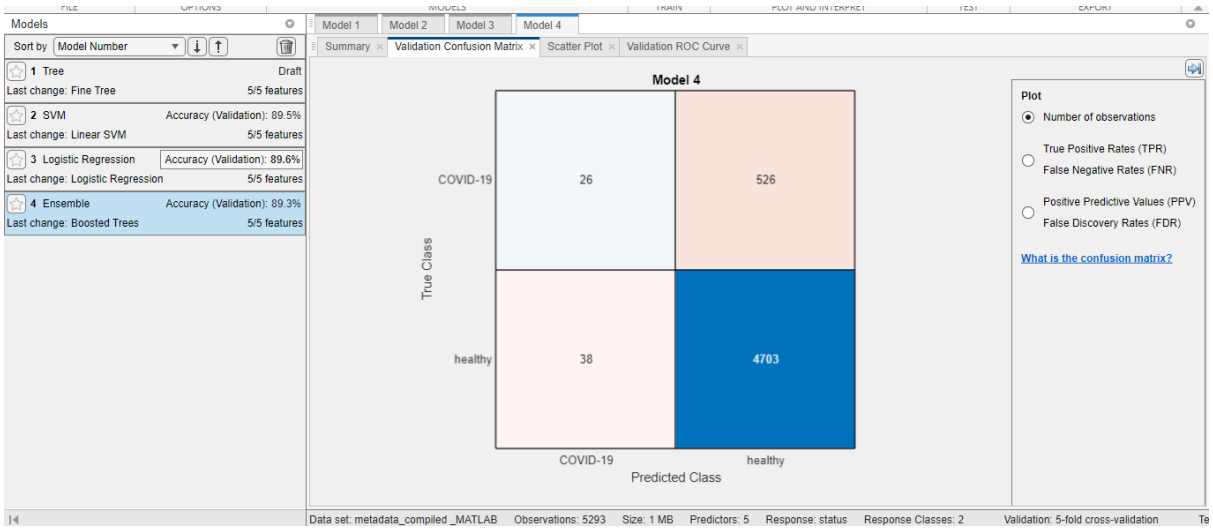

Figure 24 Ensemble confusion matrix result of no.of observations

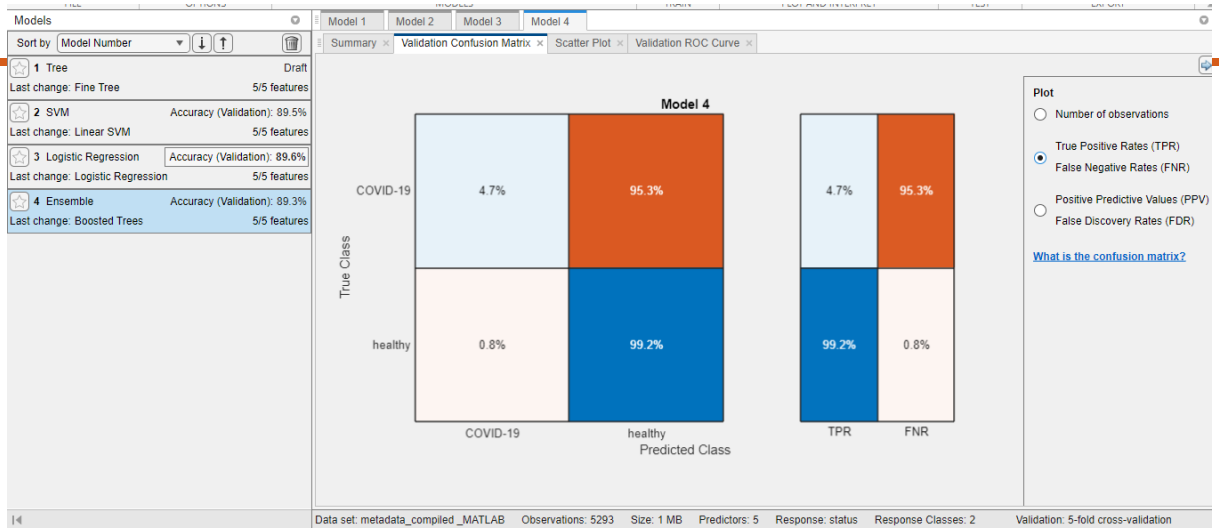

Figure 25 Ensemble confusion matrix result

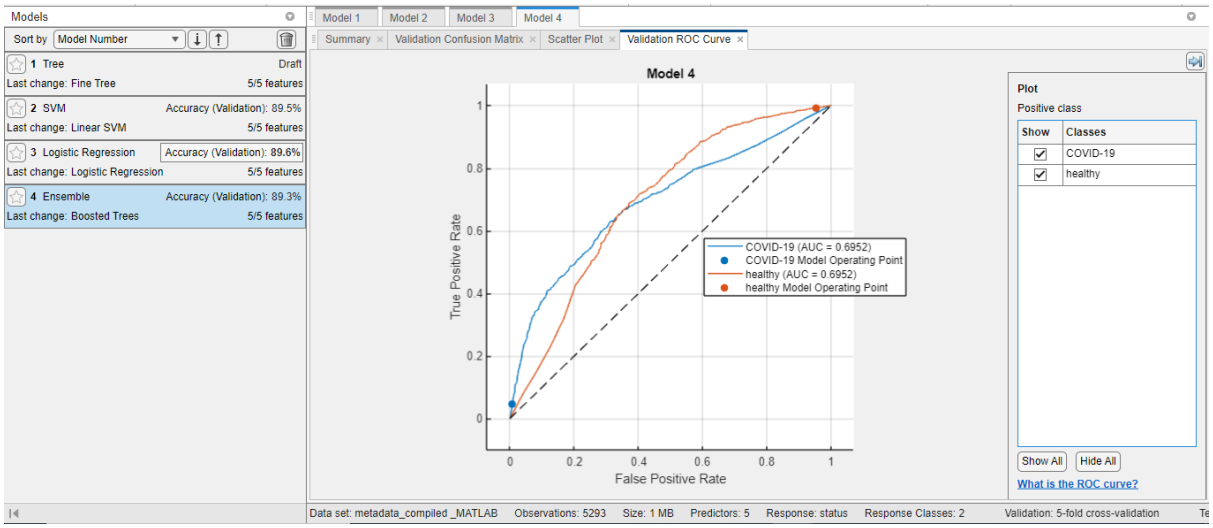

Figure 26 Ensemble ROC curve results of COVID-19 and healthy

Testing Results :

Below are the results got after testing the models on testing data in MATLAB after training the model. Tested the model in different machine learning algorithms like SVM ,logistic regression ,Ensemble. Tested data which contains covid status (49) and healthy status (137). As a response selected the status and as predictors selected the cough_detected,age,gender,respiratory_condition,fever_muscle pain.

SVM:

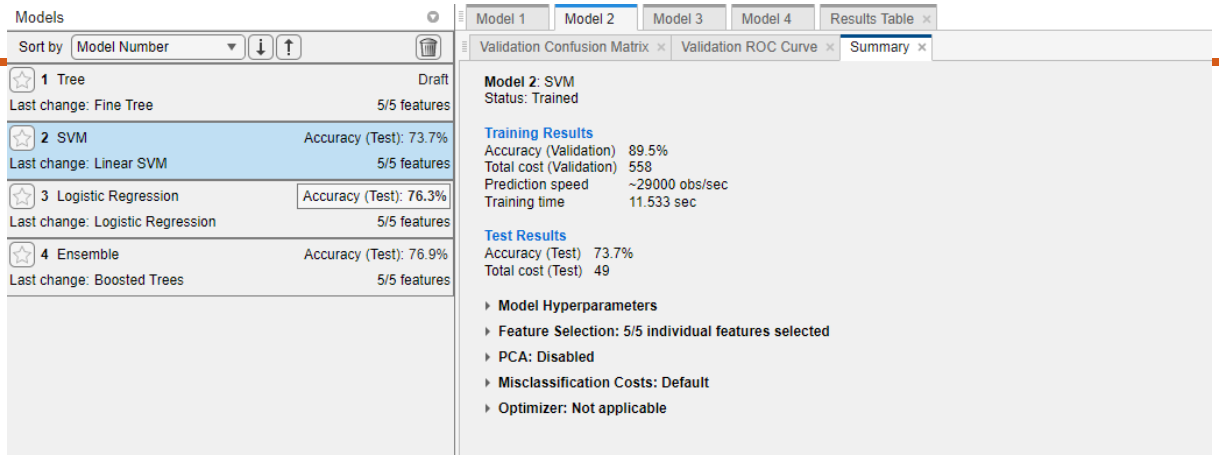

Figure 27 SVM training and testing results

Logistic Regression:

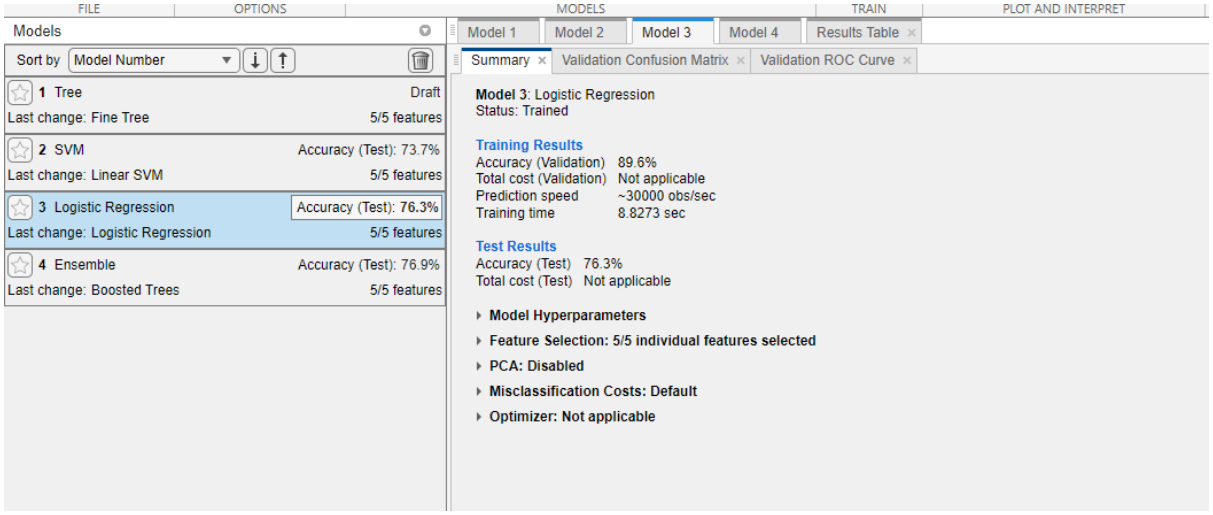

Figure 28 Logistic regression training and testing results

Ensemble:

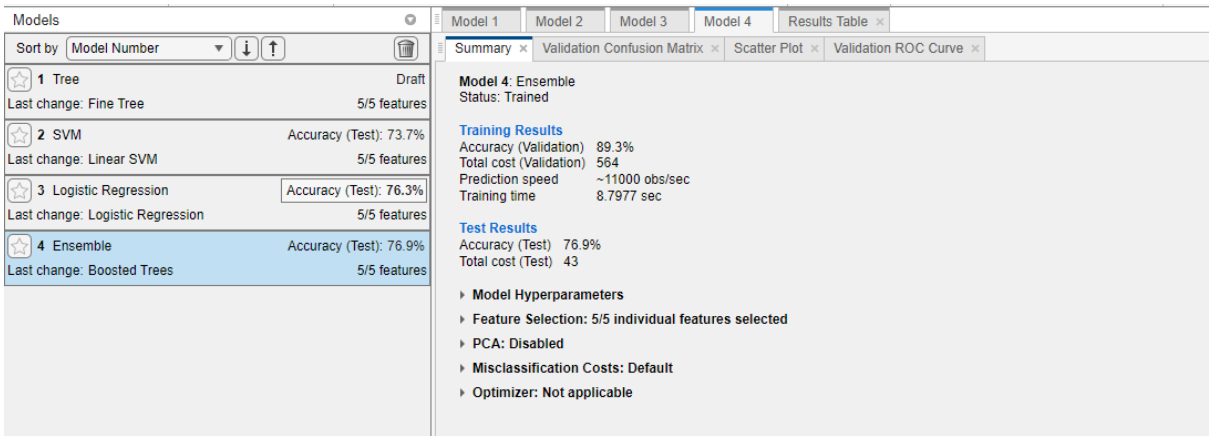

Figure 29 ensemble training and testing results

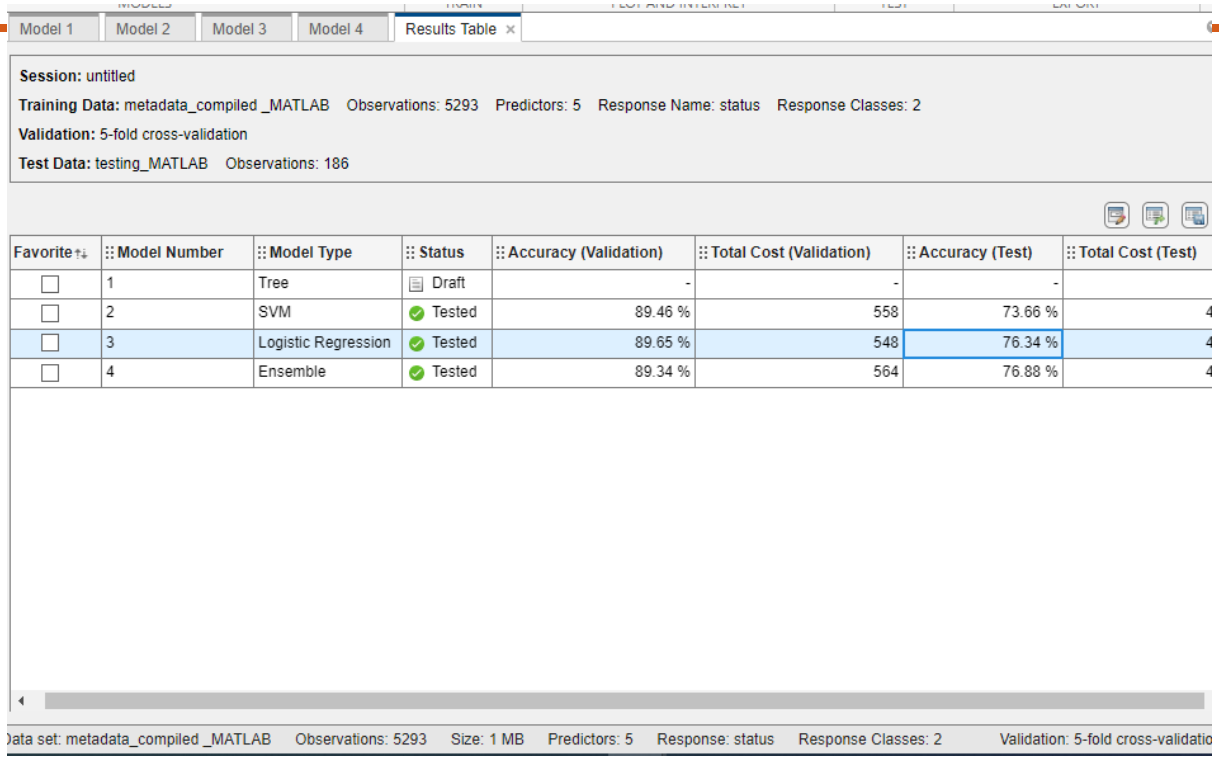

Figure 30 Testing results of SVM vs logistic regression vs Ensemble

Ensemble model got the best accuracy (validation) 89.34% and accuracy (Test) 76.34% , compared with SVM which got validation accuracy of 89.46 % and accuracy (Test) 73.66% ,logistic regression which the validation accuracy of 89.65% and Accuracy (Test) 76.34%. Among all Ensemble model got the better accuracy than SVM and linear regression.

I have used RandomForestClassifer for classifying the audio recordings. The test accuracy is 81.71% for the RandomForestClassifier implemented.

The reason for choosing RandomForest is to eliminate the overfitting that's usuaully encountered while using a DecisionTree classifier

```
print('\nBest R2 score : %.2f'%RF_grid.best_score_, ' Best Params : ', str(RF_grid.best_params_))
```

```
Best R2 score : 0.80  Best Params :  {'classifier__criterion': 'entropy', 'classifier__max_depth': 20, 'classifier__min_samples
_split': 10, 'classifier__n_estimators': 90}
```

```
print(RF_grid.best_estimator_.score(X_test,y_test))
```

```
0.8171589310829818
```

6.Conclusion and discussion:

The various learning models implemented was successful in identifying COVID-19 positive/ patients using significant audio features. Demonstrated that a crowdsourcing data collection method may provide a reliable COVID-19 detection algorithm from cough and also showed that detection algorithm continues to perform well on an external crowdsourced dataset that was gathered under slightly different conditions, in less-than-ideal surroundings, and at various stages of infection. Given that the virus signature seems to generalise, this demonstration supports the idea that COVID-19 can be accurately diagnosed via cough noises.

The datasets with comprehensive labels we chose for this project were not large enough to train the best algorithm, which represents a sample size limitation for this study. We intend to continue working toward a wider data collection. However, the majority of current approaches to data collecting do not sufficiently take into consideration diversity, including COVID-19 status labels, prior medical history, level and severity of COVID disease. Due to the fact that COVID-19 patients has different range of symptoms, which includes different combinations of anosmia, fever, asymptomatic low oxygen saturation, pneumonia, conjunctivitis, and heart injury[10]. It is questionable whether any cough-based machine learning algorithm will provide similarly accurate results for the full range of symptoms. However, a robust COVID-19 detection method from cough has widespread global applicability and can play a crucial role in limiting the disease's spread.

In Conclusion while detecting the covid using audio file random forest got the accuracy (training) got 80 % and at testing got the accuracy of 81 % .

Different Machine Learning techniques were used to train the COVID-19 classifier in MATLAB. Ensemble model got the best accuracy (validation) 89.34% and accuracy (Test) 76.34% .

7.FUTURE WORK:

As was previously mentioned, relatively few studies have been conducted to distinguish between those with COVID-19 and those who do not have COVID-19 based on cough sounds in the literature. The majority of the papers that are currently available have either been submitted for peer review or have not yet been presented at conferences. The datasets that contain Covid-19 coughing are somewhat small because this is a very new area of research. Such little datasets frequently result in overfitted deep learning models. It can be shown that the inquiry used well-known machine learning techniques, however the suggested study provided metrics that were significantly better than those produced by the other investigations. A crucial consideration to keep in mind is the system's sensitivity in the diagnosis of COVID-19 patients.

Future research can use a larger dataset to do a more effective COVID-19 diagnosis investigation based on coughing. Future research can use a variety of feature extraction and classification techniques to expand the sample size and improve classification accuracy. Identifying patterns in the metadata through analysis could be another step taken to enhance the model's performance. Biological gender, age, location, medical history, and other meta-data can be used to understand distinct patterns and determine whether they are suggestive of particular cough features. Using metadata, we may try the following analyses, for instance:
1.Are there changes in cough patterns linked to COVID-19 based on biological gender or age?
2.Does a patient's medical background influence the spread of COVID-19?
3. Are patients with asymptotic coughs distinct from those who have symptoms?
4. How do the coughing patterns of COVID-19 patients change over time, and how might longitudinal cough data from different days of an affected patient affect cough characteristics?
We can learn more about the characteristics of COVID-19 coughing by analysing these questions with a reliable dataset. Deep learning techniques can also be used to enhance the train and classification phases.

Obstacles faced:

Collecting reliable data was one of the biggest challenges encountered in this. The restrictions of utilising this data to create the model have been well discussed. Ambiguities I ran into while annotating the dataset to get it ready for the covid detection using cough sounds are yet another challenge. The concept of a cough is complicated; therefore it took me some time to understand the details and formulate a description of what a cough is in perspective of the issue at hand. Eventually, the obstacle was overcome.

Acknowledgement:

My sincere gratitude to the my supervisor during the research process for his guidance and valuable feedback. I would like to thank my family especially my husband and my 10-year-old son for their continuing support and support while I was studying.I want thanks to the university faculty members for providing fantastic assistance during my studies.

References:

1.Detection of COVID-19 Using Deep Learning Techniques and Cost Effectiveness Evaluation: A Survey by Dr. Manoj Kumar M V[1]

2.Darapaneni, N., Ranjane, S., Satya, U. S. P., Reddy, M. H., Paduri, A. R., Adhi, A. K., et al. (2020). "COVID-19 severity of pneumonia analysis using chest x rays," in 2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS) (IIT Ropar), 381–386. doi: 10.1109/ICIIS51140.2020.9342702[2]

3.El Gannour, O., Hamida, S., Cherradi, B., Raihani, A., and Moujahid, H. (2020). "Performance evaluation of transfer learning technique for automatic detection of pa- tients with COVID-19 on X-Ray images," in 2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS) (Kenitra), 1–6. doi: 10.1109/ICECOCS50124.2020.9314458[3]

4.Machine learning for detecting COVID-19 from cough sounds: An ensemble-based MCDM method Nihad KarimChowdhury, Muhammad Ashad Kabir, Md. Muhtadir Rahman[4]

5.COVID-19 cough classification using machine learning and global smartphone recordings Madhurananda Pahar, Marisa Klopper, Robin Warren, Thomas Niesler[5] https://www.sciencedirect.com/science/article/pii/S0010482521003668?via%3Dihub

6.C. Bales, M. Nabeel, C. N. John, U. Masood, H. N. Qureshi, H. Farooq, I.Posokhova, and A. Imran, "Can machine learning be used to recognize and diagnose coughs?" in 2020 International Conference on e-Health and Bioengineering (EHB). IEEE, 2020, pp. 1–4.[6]

7. C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo, "Exploring automatic diagnosis of covid-19 from crowdsourced respiratory sound data," in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 3474–3484.

8. C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo, "Exploring automatic diagnosis of covid-19 from crowdsourced respiratory sound data," in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 3474–3484.[8]

9.https://pubmed.ncbi.nlm.nih.gov/33967346[9]
10.https://towardsdatascience.com/from-mfccs-xor-gfccs-to-mfccs-and-gfccs-urban-sounds-classification-case-study-a087ac007901[10]

11.https://medium.com/prathena/the-dummys-guide-to-mfcc-aceab2450fd[2]
12.Detection of COVID-19 from speech signal using bio-inspired based cepstral features[7] Tusar Kanti Dash , Soumya Mishra , Ganapati Panda, Suresh Chandra Satapathy, https://pubmed.ncbi.nlm.nih.gov/33967346

13.https://www.rroij.com/open-access/gammatone-cepstral-coefficient-forspeaker-identification.php?aid=42795#:~:text=Taking%20as%20a%20basis%20Mel%20frequency%20cepstral%20coefficients,employing%20Gammatone%20filters%20with%20equivalent%20rectangular%20bandwidth%20bands
14.Train support vector machine (SVM) classifier for one-class and binary classification-MATLABfitcsvm,
15. https://www.sciencedirect.com/topics/computer-science/logistic-regression

16.https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

17.Artificial Intelligence–Based COVID-19 Detection Using Cough Records Alpaslan Gökcen,Bulut Karadağ1,Cengiz Riva,Ali Boyacı1

18.https://journals.asm.org/doi/10.1128/CMR.00028-20 Coronavirus Disease 2019–COVID-19 by kuldeep dharma,sharun khan

19.L. Loffredo, F. Pacella, E. Pacella, G. Tiscione, A. Oliva, and F. Violi, "Conjunctivitis and COVID-19: A meta-analysis," J. Med. Virol., vol. 92, no. 9, pp. 1413–1414, 2020, doi

20.Automatic diagnosis of COVID-19 disease using deep convolutional neural network with multi-feature channel from respiratory sound data,Cough, voice, and breath Kranthi Kumar Lella,Alphonse Pja

21. W. He, Z. Huang, Z. Wei, C. Li, and B. Guo, "Tf-yolo: An improved incremental network for real-time object detection," Applied Sciences, vol. 9, no. 16, p. 3225, 2019.

22. Imran, Ali, et al. "AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app." Informatics in Medicine Unlocked 20 (2020): 100378.

23. Laguarta, Jordi, Ferran Hueto, and Brian Subirana. "COVID-19 artificial intelligence diagnosis using only cough recordings." IEEE Open Journal of Engineering in Medicine and Biology 1 (2020): 275-281.
24. Pahar, Madhurananda, et al. "COVID-19 cough classification using machine learning and global smartphone recordings." Computers in Biology and Medicine 135 (2021): 104572.
[25] Khanzada, Amil, et al. "Challenges and opportunities in deploying COVID-19 cough AI systems." Journal of Voice 35.6 (2021): 811-812.