IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

FEATURE ANALYSIS AND DETECTION OF COVID 19 COUGH AUDIOS

[1]Dr.Jayavrinda Vrindavanam, [2]Dr.Raghunandan Srinath, [3]Hari Haran S ,[4]Gaurav Nagesh
[1]Associate Professor, [2]Principal Member of Technical Staff, Graylinx Pvt Ltd, Bangalore, [3]Student,[4]Student
Department of Electronics and Communication Engineering,
Nitte Meenakshi Institute of Technology, Bangalore, India

Abstract: This paper focuses on the contactless detection of COVID-19 patients by analyzing cough audio samples. This paper attempts to review existing methods of analyzing cough audios of lung disorders such as asthma, pneumonia and chronic cough and thereafter introduces an algorithm that can reasonably classify the COVID–19 cough audios. An audio based cough detection method can greatly mitigate the stress on frontline workers and provides an efficient and quick way to manage the time and resources of healthcare professionals. The paper has arrived at better accuracy using Random Forest classifier by selecting a series of dominant features until optimum accuracy is obtained.The proposed system has been tested on cough recordings of both COVID-19 positive and COVID-19 negative samples and the results yield reasonable classifications.

Index Terms **- COVID-19, COVID Cough, Cough Detection, Lung Disorders, Cough Sound Analysis, ML for audio analysis, ML for COVID-19 detection, ML for cough detection.**

I. INTRODUCTION

The on-going research interest in the field of audio analysis has got into focus with the emergence of COVID -19 pandemic. With the fast-emerging technologies such as Machine Learning (ML) tools and related algorithms with advanced feature extraction techniques and robust classification models, we can identify patients of COVID-19 form their cough audios. In the absence of effective remedies, early detection can be the best option to limit the spread of air bone diseases like COVID 19. Early detection would also ensure that the affected patients are quarantined well in advance and thereby the spread of the diseases can be contained; the front-line workers who deal with patients are also protected. ML algorithms play an important role in distinguishing between a COVID-19 patient and a healthy individual through the analysis of their respective cough patterns. Cough is associated with a characteristic sound and in this paper the pattern of COVID-19 cough is identified by performing certain feature extraction and grid search techniques. The paper proposes to analyse the frequency, duration, and image pattern of the cough waveform whereas the alternative approach has been through the chemical analysis of conducting cheek swab, nose swab, and blood test which are tedious in nature and the results take up to 2 days for Reverse Transcription Polymerase Chain Reaction (RT-CRT) tests. Antigen tests can provide the result in a matter of 30 minutes, the results at times may not be accurate, still induces strain on the chemical labs and also generates chemical and toxic wastes. The alternative is opting for a digital testing process. In this paper, we are proposing to analyse the cough audios of patients into frames and then analyze the waveforms based on different parameters to extract different features and rank them to obtain dominant features to supply to the classifier for classifying the audio into that of a COVID-19 patient or a non-COVID 19 person using Random Forest classifier. As part of the process, the paper proposes to provide the five most important features to these three classifiers to obtain a quick result, and subsequently add five more features and perform classification. This process is iterated for number of features for numbering 10, 15 and 25 and the most the most appropriate feature set is selected for real-world implementation of the COVID-19 cough detection.

II. LITERATURE SURVEY

Since the key focus of the paper is to identify COVID-19 patients by their cough audios, this review attempts to provide an overview of the prevailing key studies in the area of cough detection and identifying diseases based on the frequency, duration, and intensity of cough audio samples. Keeping in view the trends in technological advancements specific to the focus of this research, the paper has reviewed a few similar studies that have used cough data analysis for different lung-based diseases by developing a classifier on features extracted from cough audios of the patients [2] [5]. In a study carried out by of S. Matos Et. al [2] measuring the intensity and frequency of the coughs of patients can be used to identify asthma. However, this detection was done by attaching microphones to the chest of patients and thus obtaining an average detection rate of 82%. Using Hidden Markov's Model (HMM) as the classifier, the authors were able to identify asthma patients with an accuracy of 77% [2]. In another study [1], the focus was on classifying dry cough and wet cough. The paper explained different feature extracting techniques that can be performed into classifying cough as dry or wet. The study converted .mp3 files into .wav files to extract frequency domain features. Energy envelope peak detection and power ratio estimation were extracted. The authors divided the cough waveform into three phases: Initial opening burst, noisy airflow, and glottal closure. These three phases were different for

dry and wet cough thus making it easier for differentiation [1]. Sinem[4]et. al. made use of auscultation (listening to lung sounds using stethoscope) using ML algorithm to identify crackling sounds for detection of cough as crackling sounds are discontinuous in nature and shorter than 100ms and caused by pressure balancing and pressure change due to sudden opening of closed airway in lungs. The authors have developed a classification system to distinguish between normal and abnormal lung sounds using Artificial Neural Networks (ANN) and Support Vector Machines (SVM) [4]. In another study [5]attempted was made to identify pneumonia based on the mathematical analysis of cough sounds with cough features inspired by wavelet-based crackle detection work in lung sound analysis. By extracting 30 cough features from each sample and using Logistic Regression Model the authors were able to obtain 94% and 88% sensitivity and specificity by combining wavelets with other features [5].

In a paper by Jesus Monge Et. al [6], the authors have the proposed an approach that computes short-term features in relevant frequency bands; the most meaningful features were then selected and combined in a high-level representation to perform robust cough detection in noisy conditions. A short-term spectral feature set was separately computed in five predefined frequency bands. Feature selection and combination tasks then applied to make the short-term feature set robust enough in different noisy scenarios. High-level data representation was achieved by computing the mean and standard deviation of short-term descriptors in 300 ms long-term frames. The classification system using Support Vector Machines (SVM) achieved 92.71% sensitivity, 88.58% specificity, and 90.69% Area Under Receiver Operating Characteristic (ROC) curve (AUC) [6]. Another paper attemped to identified cough from acoustic signals. The authors used a classification model of Logistic Regression based on 4 features. Based on the 4 fundamental properties of cough: Widespread Spectrum, Low Tone Prominence, Sudden Burst of Energy, and Short Duration, the study classified audio sample as a cough or not a cough using LRM and noted down sensitivity (SE) and Specificity (SP) for different model like Deep Neural Networks, Artificial Neural Networks and K-Nearest Neighbour. A combination of Leicester cough monitor and Hull automated cough counter using Logistic Regression Model acted as the required classifier to achieve an average sensitivity achieved was 86.78% [8] . In another relevant study [9], the cough sound signals of the patients were predicted/classified into different respiratory disorders using Support Vector Machine (SVM) classifier with 3 features extracted. SVM was used as a hyperplane which divided the data into different classes and thus capable of classification of cough audio into different diseases. SVM classifier yielded an accuracy of 98.9% with True Positive Rate (TPR) ranged from 94% to 100%, False Negative Rate (FNR) was from 5% to 6% [9].

In the process of detection of COVID-19 patients by analyzing the cough audios Chloë Brown Et. al University of Cambridge have done crucial work in collecting cough audio samples 7,000 users with 200 users tested positive for COVID-19 and shared the dataset with us upon request. The reserachers used CNN to identify cough audios along with a combination of Support Vector Machines (SVM) wit Radial Basis Function (RBF) and Data Argumentation as classifier. Area Under Curve (AUC) for COVID-19 cough vs. Non-COVID-19 cough using VGGish was 82% [10]. In another study on COVID-19 Cough, data extracted from web-based application was classified using Random Forest Classifier with 30 trees based on 9 features extracted from cough audios and thus obtaining an accuracy of 66.74% [11].

III. RESEARCH METHODOLOGY

3.1Data and Sources of Data

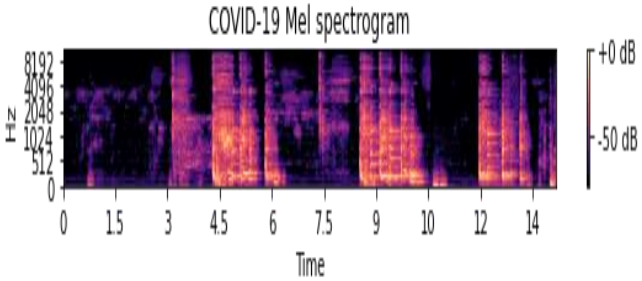
We have obtained the dataset from University of Cambridge,who had collected cough recordings of 10,000 samples from 7,000 individuals, out of which 2,000 were tested positive for COVID-19 and the negative recordings were collected from the countries were the virus was not spread at the time of recording out of which we were provided with 104 cough audio recordings at 44KHzwith 54 COVID-19 positive patients and 50 healthy individuals. We have further added 18 healthy cough audio recordings from Free sound database [14] and 46 healthy cough audio recordings from Coswara database of Indian Institute of Science, Bengaluru, India increasing the overall healthy audio recordings to 96 and the overall total to 150 cough audio recordings.

3.2 Feature Analysis

We have plotted the required features in order to visually analyze and understand which features can cause significant impact on the ML classification models. We have considered 64 features under 8 categories which are decisive to determine vocal features and detect cough audios for classification; the categories considered are listed below:

3.2.1 MFCC

The Mel Frequency Cepstral Coefficients(MFCC) feature extraction technique basically includes windowing the signal, applying the DFT, taking the log of the magnitude, and then warping the frequencies on a Mel scale, followed by applying the inverse DCT. We can observe a visual comparison of Mel spectrogram betweenCovid-19 and healthy individuals in figure 1.



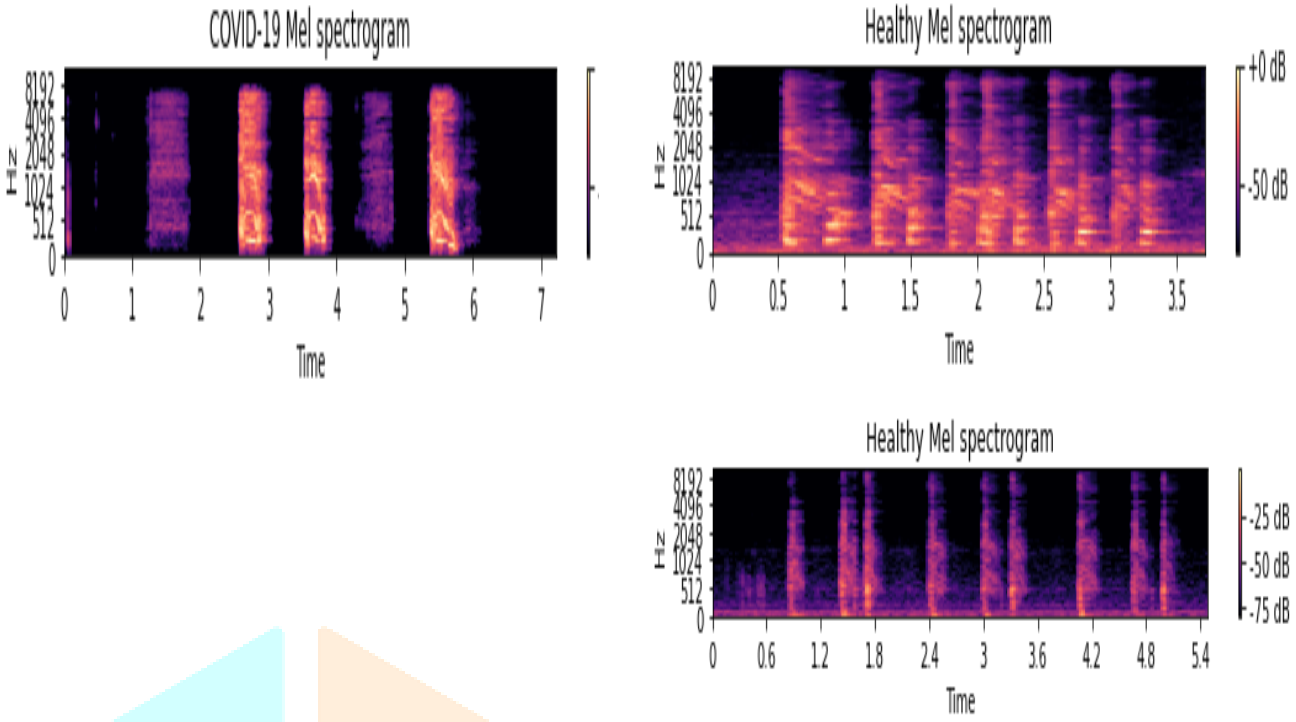


Fig 1: Comparison of MFCC between COVID-19 Positive and healthy individuals.

3.2.2 Zero Crossing Rate

Zero-crossing rate is a measure of the number of times in each time interval/frame the amplitude of the speech signals passes through a value of zero. As seen in the figure 2, we can visually compare Zero Crossing Rate between Covid-19 and healthy individuals.

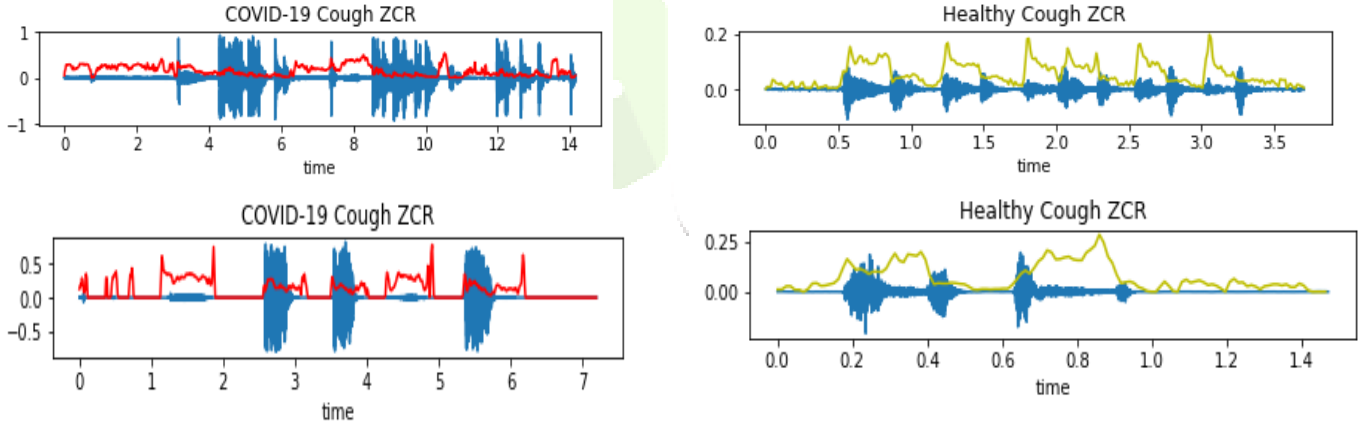


Fig 2: Comparison of Zero Crossing Rate between COVID-19 Positive and healthy individuals.

3.2.3 Root Mean Square Energy

The RMS value of continuous-time waveform is the square root of the arithmetic mean of the squares of the values, or the square of the function that defines the continuous waveform. The Figure 3 shows the visual comparison of Root Mean Square Energy between Covid-19 and healthy individuals.

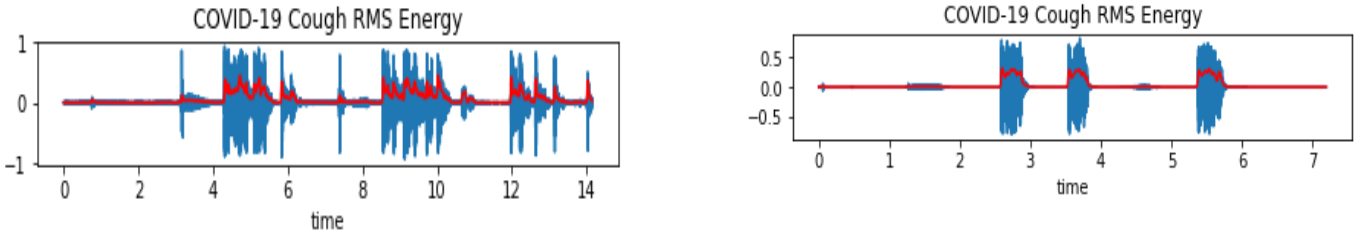


Fig 3: Comparison of Root Mean Square Energy between COVID-19 Positive and healthy individuals.

3.2.4 Power Spectral Density

3.3 A Power Spectral Density (PSD) is the measure of signal's power content versus frequency. PSD helps to ensure that random data can be overlaid and compared independently of the spectral resolution used to measure the data. A visual comparison of Power Spectral Density between Covid-19 and healthy individuals is given in Figure 4.

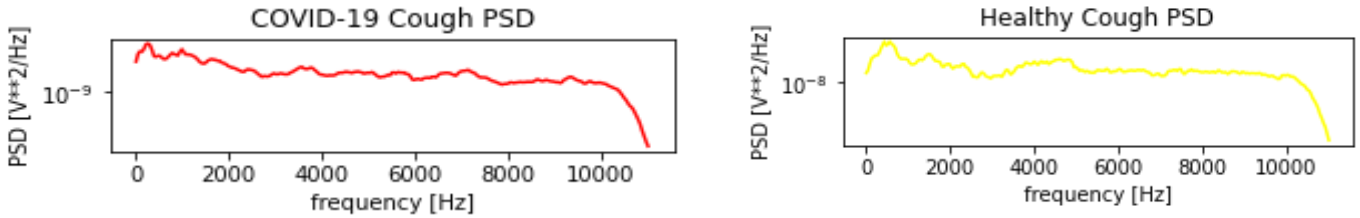


Fig 4: Comparison of Power Spectral Density between COVID-19 Positive and healthy individuals.

3.3.1 Dominant Frequency

Dominant frequency is strictly defined as the frequency with the largest amplitude on a spectrum. Dominant frequency is found by decomposing the electrograms into a finite number of sinusoidal constituents and finding the one with the highest magnitude. (Figure 5).

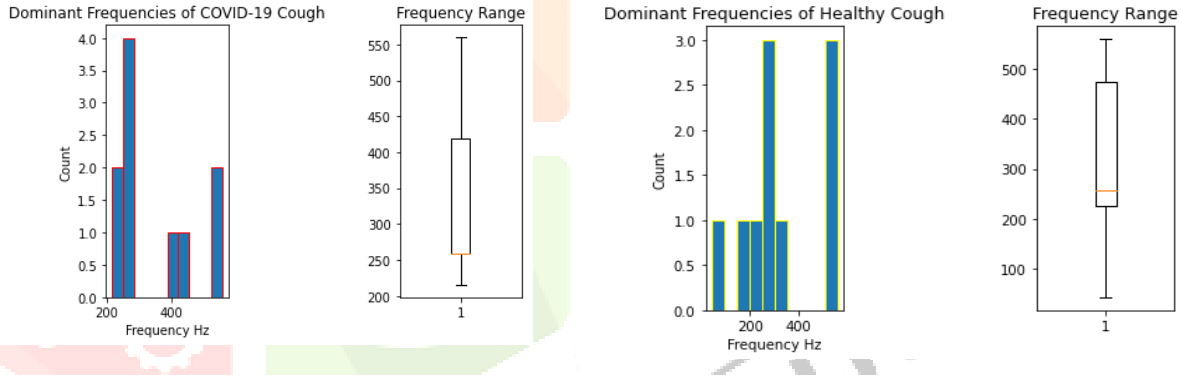


Fig 5: Comparison of Dominant Frequency between COVID-19 Positive and healthy individuals.

3.3.2 Crest Factor

Crest factor indicates how extreme the peaks are in a waveform. Crest factor is a parameter of a waveform showing the ratio of peak values to the effective value. The visual comparison is provided in Figure 6.

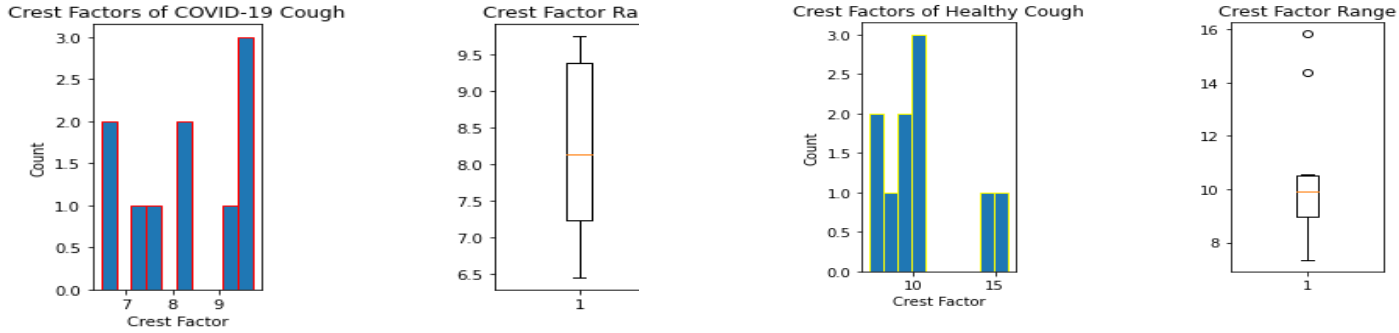


Fig 6: Comparison of Crest Factor between COVID-19 Positive and healthy individuals.

3.3.3 Spectral Centroid

The spectral centroid helps in showing at which frequency the energy of a spectrum is centred upon. Spectral centroid is a measure used in digital signal processing to characterize a spectrum and indicates the location of the centre of mass of the spectrum. A visual comparison of Centroid between Covid-19 and healthy individuals is shown in figure 7.

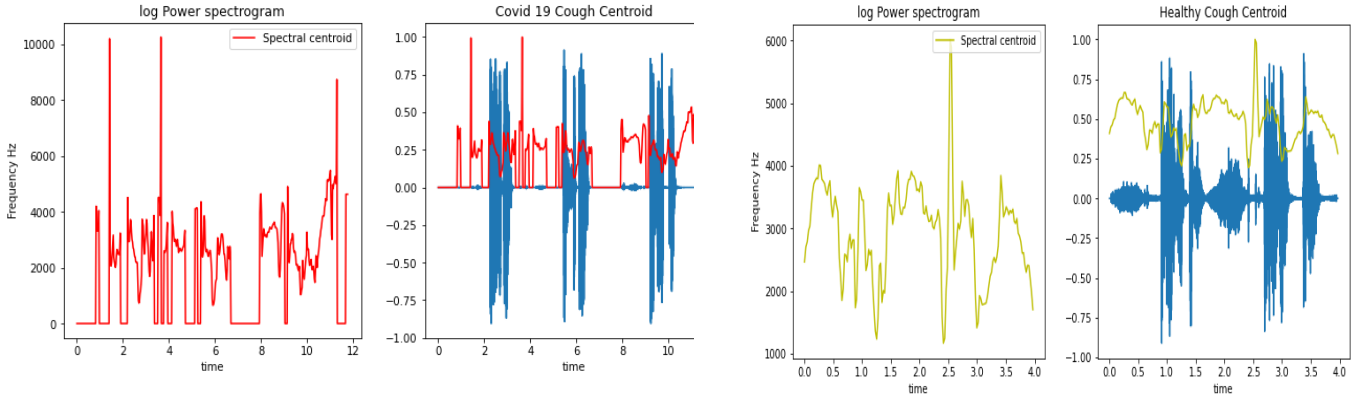


Fig 7: Comparison of Spectral Centroid between COVID-19 Positive and healthy individuals.

3.3.4 Spectral roll-off

Spectral roll-off is the frequency below which a specified percentage of the total spectral energy lies. A visual comparison of Spectral roll-off between Covid-19 and healthy individuals is shown in figure 8.

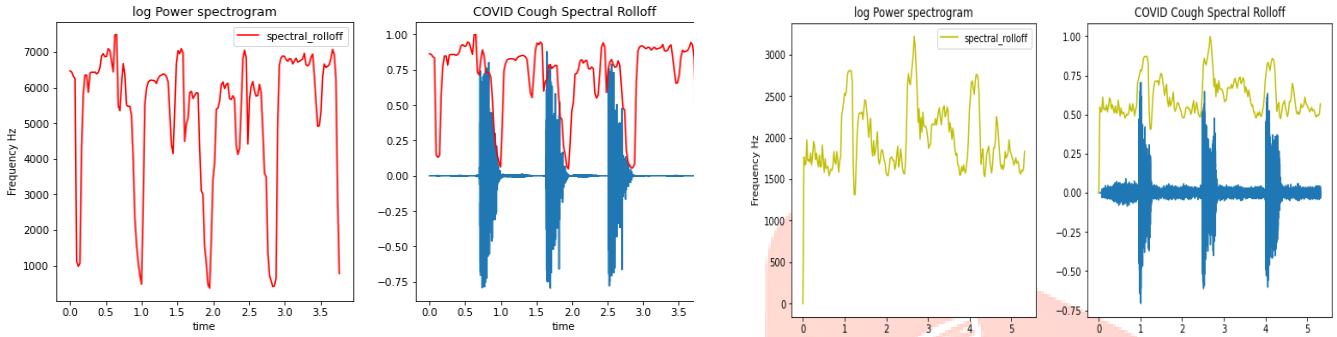


Fig 8: Comparison of Spectral roll-off between COVID-19 Positive and healthy individuals.

3.3 Theoretical framework

The plotting of the features show visual distinction in figures indicating classificatory power of the data. The classification system has been implemented using Python (Jupyter Notebook). The paper has used Random Forest to classify the input cough audios into that of a COVID-19 positive patient and COVID-19 negative person (Healthy person) the model is rated on accuracy, precision, recall and F1-score.

The following are the key steps involved in the processing:-

3..3.1 Access and resample the cough audio files to 22KHz and convert stereo to mono channel.

3..3.2 Extract key features from the audio files such as MFCC (first 12 co-efficient), Zero Crossing Rate, Root Mean SquareEnergy, Dominant Frequency, Crest Factor, Power Spectral Density, Spectral Bandwidth, Spectral Centroid, Linear Predictive Coding Co-efficient, first 4 formants, Log Energy and Statical features and determine the dominant features by using Weight of evidence, Random Forest Feature importance, Recursive Feature Elimination,Chi Square value, Extra Tree Classifier, L1 feature selection and ranking the features. The figures (1 to 8) shows the pattern of different features of healthy and covid 19 persons.

3..3.3 Normalize the features for easier computation to feed them to classifier.

3..3.4 The five most dominant features are selectedusing XuniVerse module which is open sourced and freely available on GitHub [15] and passed to Random Forest classifier and the accuracy, recall, pression and F1-score are calculated for this initial classification purposes as the processing time is comparatively much lower.

3..3.5 After the initial classification, we pass the ten most dominant features to Random Forest classifier to obtain a precise result as the precision and accuracy increases at the cost of computation time.
This process is iterated for number of features 15 and 25 to obtain a highly specific classifier.

3..3.6 The ML model is trained by using k-fold cross validation so that each example is tested once and classified into COVID-19 positive and COVID-19 negative cough audios.

3..3.7 Compare the decision parameters for different set of dominant features to find out the set of features to be paired up with the classifiers for implementation.

IV. RESULTS AND DISCUSSION

The classification model is tabulated based on the parameter's accuracy, recall, precision, and F1-score for different set of features:

5 Features:

	Variable_Name	Information_Value	Random_Forest	Recursive_Feature_Elimination	Extra_Trees	Chi_Square	L_One	Votes
0	spectral_flatness_Q1	1	1	1	1	1	1	6
7	I13	1	1	1	1	1	1	6
20	m2	1	1	1	1	1	1	6
16	spectral_bandwidth_p2_Q3	1	1	1	1	1	1	6
1	spectral_flatness_Q2	1	1	1	1	1	1	6

Fig 9: First 5 Features extracted based on votes.

Table 1: Comparison between 5 features,10 features,15 features and25 feature.

Features	Accuracy	Recall	F1-Score	Precision	Wall Time
5 Features	80.50%	70.00%	71.50%	75.10%	72.10ms
10 Features	85.20%	73.81%	78.15%	84.90%	106.00ms
15 Features	85.20%	73.81%	78.07%	84.73%	133.00ms
25 Features	83.20%	73.80%	76.10%	80.84%	138.00ms

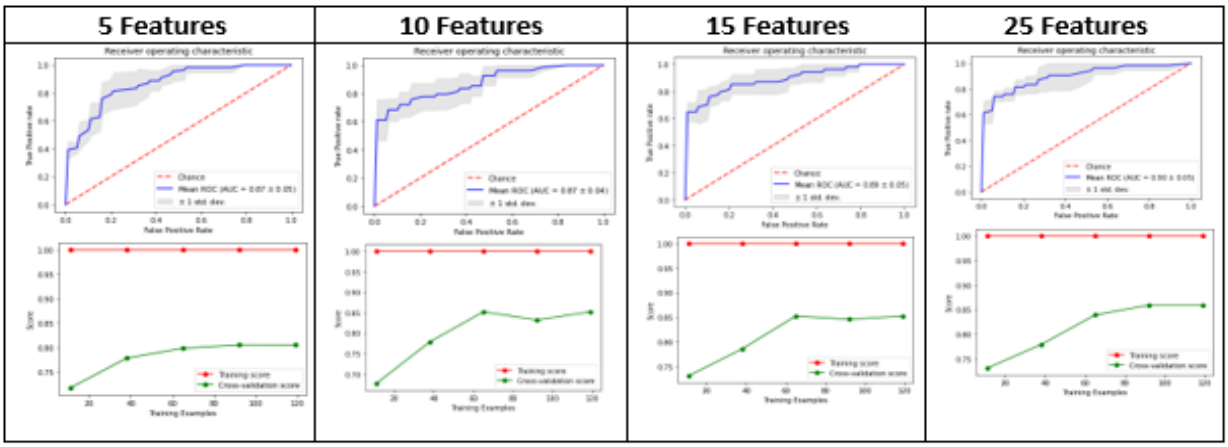


Fig 10: The mean Receiver Operating Characteristics (ROC) along with its Area Under the Curve (AUC) acrossvarious folds and the learning curves of all the three models are tabulated.

From the above table and Reciever Operating Characteristics, we can come to the conclusion that the feasible number of features to be used along with Random Forest classifier is 10, as both 15 features and 25 features result in the same range of accuracy and recall rate while considerably increasing the wall time and when only 5 features are considered the accuracy and recall rate in much lower.

V. CONCLUSION AND FUTURE WORK

The project has presented a novel approach in identifying COVID-19 patients using a contactless procedure of analyzing the cough audio recordings of a person. The paper demonstrates ML classification model of Random Forest along with 5, 10, 15 and 25 features and determines the optimal number of features to be used along with the model for implementation based on parameters such as computational complexities, accuracy and recall. The work-initiated forms part of a larger project of ML based classification model which includes developing either hardware or software applications for easy detection. Extracting robust features for classification and implementing the algorithm into an integrated circuit which yields into a real-world classification model would be taken up as a future work.

Acknowledgements

The authors gratefully acknowledge Cambridge University and Professor Cecilia Mascolo, Department of Computer Science and Technology for making available the COVID 19 sounds data for enabling us to perform this research work, without their generous support, this analysis would not have been possible. We express our sincere thanks to them. We also gratefully acknowledge the management of NITTE Institute of Technology for extending necessary support in the student research projects.

REFERENCES

[1] Hanieh Chatrzarrin , Amaya Arcelus , Rafik Goubran and Frank Knoefel.2011.Feature extraction for the differentiation of dry and wet cough sounds. 2011 IEEE International Symposium on Medical Measurements and Applications, INSPEC Accession Number: 12138224, DOI: 10.1109/MeMeA.2011.5966670.

[2] S. Matos, S.S. Birring, I.D. Pavord and H. Evans.2006 .Detection of cough signals in continuous audio recordings using hidden Markov models" Published in: IEEE Transactions on Biomedical Engineering, INSPEC Accession Number: 8927242, DOI: 10.1109/TBME.2006.873548, Volume:53: 1078 to 1083.

[3] Payam Moradshahi, Hanieh Chatrzarrin and Rafik Goubran.2013.Cough sound discrimination in noisy environments using microphone array. IEEE International Instrumentation and Measurement Technology Conference (I2MTC), INSPEC Accession Number: 13662626, DOI: 10.1109/I2MTC.2013.6555454.

[4] Sinem Uysal, Hüsamettin Uysal, Bülent Bolat and Tülay Yıldırım.2014.Classification of normal and abnormal lung sounds using wavelet coefficients" Published in: 2014 22nd Signal Processing and Communications Applications Conference (SIU), INSPEC Accession Number: 14381701, DOI: 10.1109/SIU.2014.6830685.

[5] Keegan Kosasih, Udantha R. Abeyratne, Vinayak Swarnkar and Rina Triasih.2014.Wavelet Augmented Cough Analysis for Rapid Childhood Pneumonia Diagnosis. IEEE Transactions on Biomedical Engineering, INSPEC Accession Number: 15000297, DOI: 10.1109/TBME.2014.2381214, Volume: 62: 1185 to 1194.

[6] Jesús Monge-Álvarez , Carlos Hoyos-Barceló , Luis Miguel San-José-Revuelta and Pablo Casaseca-de-la-Higuera.2019.A Machine Hearing System for Robust Cough Detection Based on a High-Level Representation of Band-Specific Audio Features. IEEE Transactions on Biomedical Engineering, INSPEC Accession Number: 18831922, DOI: 10.1109/TBME.2018.2888998, Volume: 66: 2319 to 2330.

[7] Ahmad Taquee, Vikrant Bhateja, Adya Shankar and Agam Srivastava.2018.Combination of Wavelets and Hard Thresholding for Analysis of Cough Signals.Second World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), INSPEC Accession Number: 18380559, DOI: 10.1109/WorldS4.2018.8611597.

[8] Renard Xaviero Adhi Pramono , Syed Anas Imtiaz and Esther Rodriguez-Villegas.2019.Automatic Identification of Cough Events from Acoustic Signals. 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), INSPEC Accession Number: 19126595, DOI: 10.1109/EMBC.2019.8856420.

[9] Vikrant Bhateja, Ahmad Taquee and Dilip Kumar Sharma.2019.Pre-Processing and Classification of Cough Sounds in Noisy Environment using SVM.4th International Conference on Information Systems and Computer Networks (ISCON), INSPEC Accession Number: 19454818, DOI: 10.1109/ISCON47742.2019.9036277.

[10] Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Jing Han, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta and Cecilia Mascolo.Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data.University of Cambridge.

[11] Neeraj Sharma, Prashant Krishnan, Rohit Kumar, Shreyas Ramoji, Srikanth Raj Chetupalli, Nirmala R., Prasanta Kumar Ghosh, and Sriram Ganapathy.Coswara - A Database of Breathing, Cough, and Voice Sounds for COVID-19Diagnosis. Indian institute of science

[12] Vinayak Swarnkar, Udantha R. Abeyratne, Anne B. Chang, Yusuf A. Amrulloh, Amalia Setyati, and Rina Triasih.2013.Automatic Identification of Wet and Dry Cough in Pediatric Patients with Respiratory Diseases. Annals of Biomedical Engineering.

[13] McFee, Brian, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto.2015.librosa: Audio and music signal analysis in python.In Proceedings of the 14th python in science conference: 18-25.

[14] The Freesound database website. [Online].Available: https://freesound.org/

[15] XuniVerse Package. [Online]. Available: https://github.com/Sundar0989/XuniVerse