Proceedings of the Fifth International Conference on Intelligent Computing and Control Systems (ICICCS 2021)
IEEE Xplore Part Number: CFP21K74-ART; ISBN: 978-0-7381-1327-2

PRELIMINARY DIAGNOSIS OF COVID-19 BASED ON COUGH SOUNDS USING MACHINE LEARNING ALGORITHMS

Arup Anupam
Deptt. of E&I
NIT Silchar
arupanupam@gmail.com

N Jagan Mohan
Deptt. of ECE
NIT Silchar
jaganmohan427@gmail.com

Sudarsan Sahoo
Deptt. of E&I
NIT Silchar
sudarsan@ei.nits.ac.in

Sudipta Chakraborty
Dept.ofE&I
NIT Silchar
sudipta@ei.nits.ac.in

2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS) | 978-1-6654-1272-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICICCS51141.2021.9432324

Abstract— **The proposed work is focused on COVID-19 classification of cough sounds based on machine learning which is used to differentiate COVID-19 coughs from non COVID-19 and healthy coughs. It follows a non-contact based screening test which is very easy to apply being non-invasive and simply carried out within the boundaries of home so that the medical testing centers are not over flooded with patients and there is an overwhelming pressure because of maintenance of those patients with shortage of adequate infrastructure facilities. The dataset used in this study has been derived from the Coswara database which comprises of around 160 infected and 480 healthy individuals. Therefore, Artificial Intelligence based machine learning classifiers were used as an alternative means of diagnosis. Logistic regression (LR), K- Nearest neighbor (KNN), support vector machines (SVM), decision tree algorithms were used as classifiers in the proposed work. The results of this study show that the SVM classifier turned out to be the best in comparing among the COVID-19 and non COVID- 19 coughs with area under receiver operating characteristic curve (ROC) of 0.98.The novelty in the proposed work includes the collection of dry cough samples which would aid in preliminary diagnosis of the infection. This form of classification can also be implemented in a smart phone after performance evaluation from medical authorities.**

Keywords— Cough sounds, COVID-19, Diagnosis, Classification, Machine learning.

1. INTRODUCTION

COVID-19 emerged in first quarter of 2020 and was labelled pandemic by WHO has created havoc across the globe and is caused by (SARS-CoV2), a novel variant although somewhat similar to the family of coronaviruses which caused SARS and MERS but the level of contagiousness is far too high [1]. In the initial phase of the infection, a person most commonly exhibit symptoms of fever, dry cough, loss of smell or taste [2]. However gradually several other symptoms like fatigue, muscle stiffness, shortness of breath appear within the patient [3]. This pandemic has caused the healthcare system to be overloaded in terms of patient testing and monitoring as the testing kit and PPE (personal protective equipment) used by the medical diagnostic bodies are limited in number and in no way can match the alarming rate of infection spread[4],[5]. In the beginning phase of COVID-19, when the virus has not yet entered into the tracheal cavity (respiratory tract which leads to the bronchioles of lungs), dry cough prevails a result of hypersensitivity reaction which is an outcome when the virus starts binding to the ACE-2 receptors on body cells. There is no production of mucus or sputum in this stage. But once the virus enters the lungs which usually happen within 1 week of infection, wet cough is experienced due to accumulation of WBCs and other matter and there is production of mucus or sputum. Hence there is an urgent need to employ this pre-screening test for people experiencing persistent dry cough and are willing to self-isolate them and take necessary precautionary measures to help control the spread of this infection.

1.1 Pre-Screening Approach towards COVID-19 based on Cough sounds

a) Even though cough is one of the predominant symptoms, it does not manifest in certain COVID-19 patients [6].But studies have revealed that coughing seems to be one of the major factors in spreading of COVID-19 at a larger scale [7]. Droplets from the cough which contains the virus may land on surfaces where it could last for long time and may be potential factor for spreading the disease further. Hence cough based testing even though quite insensitive when compared to clinical testing can help in reducing the spread as a person who is not exhibiting cough symptom may not actually spreading the disease to an extent a person with virus and cough symptom does.

b) Assessing the temperature of patients has also emerged as a pre-screening tool recently. However factors causing fever due to medical ailments other than COVID-19 are numerous in comparison to factors causing cough due to medical illness other than COVID-19.Analysis made by researchers has revealed that cough contains certain unique COVID-19 features which are also implicit with non-spontaneous cough, thereby indicating that cough based pre-screening mechanism can be implemented by forcibly allowing the person to cough if that is not spontaneous.

c) The samples derived from COVID-19 patients include basically cough sound recordings which are both symptomatic and non-spontaneous as well. This is done to check whether a person not

exhibiting cough as a symptom might have been infected. Also cough samples from completely healthy individuals with no past records of medical illness are generally collected which has been classified as normal cough. This cough is analyzed in the study in order to differentiate it from the COVID-19 cough.

d) CT scan images of chest have been widely utilized in detection and screening of COVID-19 [8], [9]. Also, Speech signals work as a biomarker for analyzing the health status of a COVID-19 patient. Therefore, a health state based detection system can be worked out to observe the sleep-quality, fatigue, depression and anxiety after suffering from the illness and losing their near and dear ones [10].

Cough can be a symptom of many underlying diseases. In fact, it is possible to assess the type of illness through test of the auditory features using multiple classifiers. In this paper, therefore a preliminary screening and diagnosis of COVID-19 is proposed mainly dealing with the cough sound analysis of different extracted features using several machine learning classifier algorithms.

2. METHODOLOGY

The methodology which is depicted in Fig 1 encompasses sections namely data collection, data processing, feature extraction and classification.

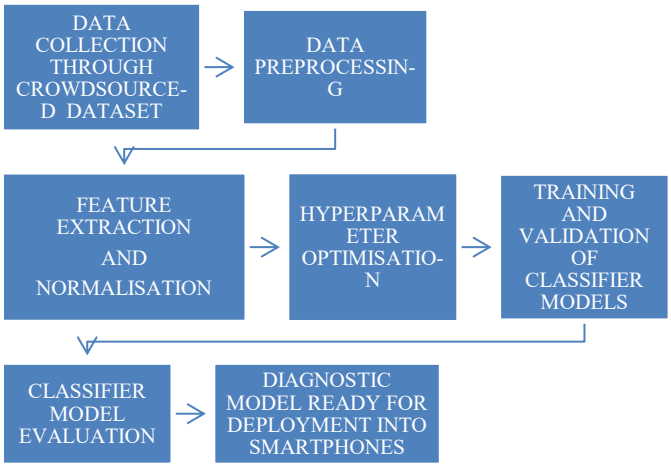

Fig 1. Block Diagram of Methodology

2.1 Data Collection

The cough samples collected for the study have been taken from Coswara project which aims at developing diagnostic tool for COVID-19 based on cough, breathing and voice sounds[11].Around 336 normal healthy samples were used for training and 144 samples were utilized for testing purposes. Similarly, 112 infected samples were considered for training while remaining 48 samples were utilized for testing purposes.The access to this web based data collection is available in the website (https://coswara.iisc.ac.in). Other than cough samples several other parameters are also recorded for the patient like age, gender, current health status, geographical location, pre-existing medical conditions. Health status includes 'normal', 'abnormal' where the patients currently not infected with the virus are categorised as normal while patients already infected with the virus are categorised as abnormal. The audio recordings are done at sample frequency of 44.1 KHz and samples are taken from people of different continents with maximum contribution from Asia and all other continents excluding Africa as illustrated in Fig 4.

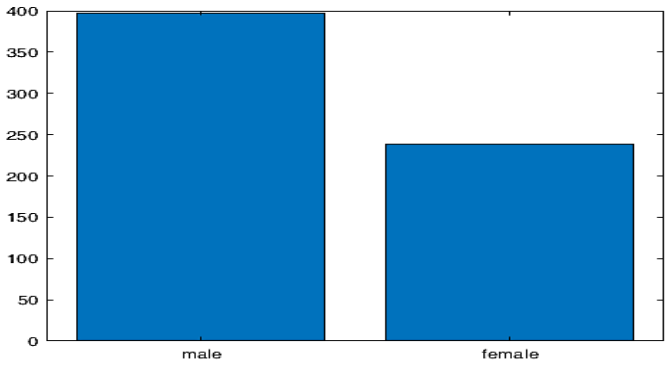

Fig 2. Number of male and female subjects

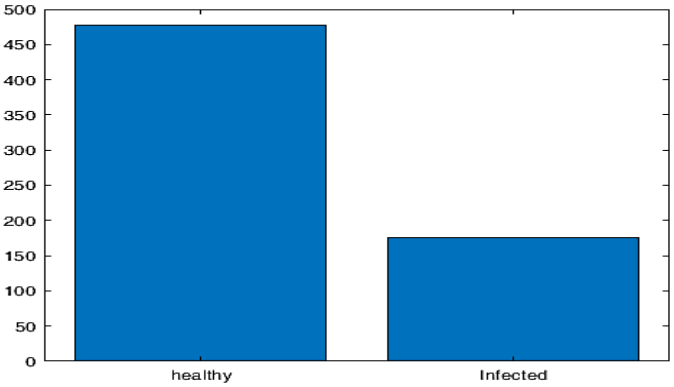

Fig 3. Number of healthy and COVID-19 positive subjects

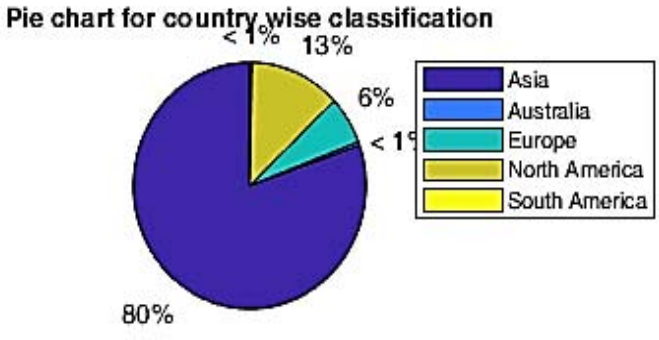

Fig 4. Pie chart of country wise classification

2.2 Data Pre-Processing

The raw audio recordings were normalized in their amplitudes and the silence period from leading and trailing ends were removed from the audio sample using Wavelet Toolbox feature available in MATLAB. The original raw audio sample, the preprocessed audio sample and the spectrogram of the contaminated cough signal is shown in Fig 5, Fig 6 and Fig 7 respectively.

These audio files have a uniform sampling rate of 44.1 KHz and the total duration of healthy and COVID-19 subjects are 0.45 hours and 8.45 minutes respectively. The average length for these two categories of subjects is 3.375 sec and 3.16 sec respectively as illustrated in Table 1.

TABLE 1 Time duration of audio recordings

	Total no of subjects	Total Length	Average length
Coswara COVID Positive	480	0.45 hours	3.375 sec
Coswara Healthy	160	8.45 minutes	3.16 sec
Coswara Total	640	0.59 hours	3.323 sec

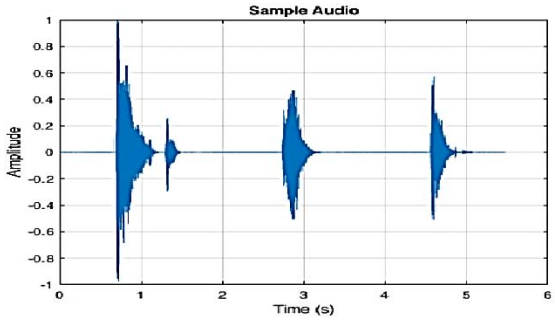

Fig 5. An original COVID-19 cough recording

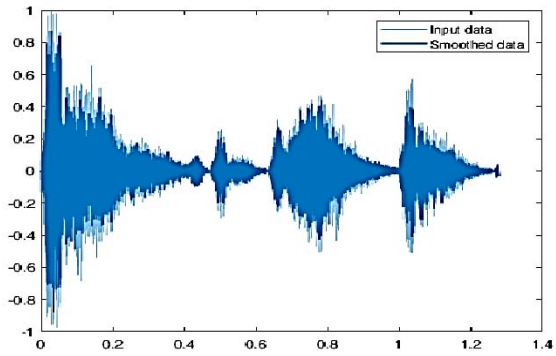

Fig 6. The Processed COVID-19 cough recording

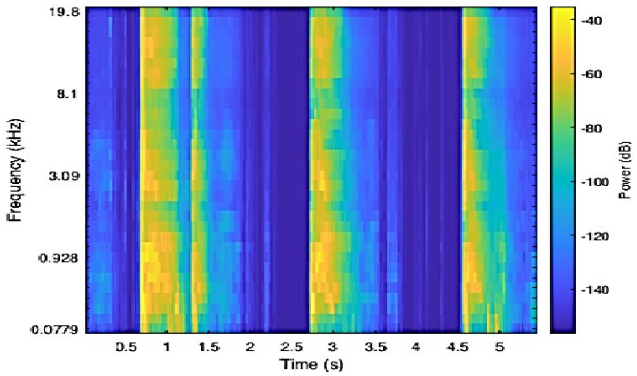

Fig 7. Spectrogram of contaminated cough signal

2.3 Feature Extraction

The Feature extraction process comprises of dividing the processed audio recording into segments which contains chunks of information in the form of frames which are further exploited to derive necessary features which shall be utilized in model training. Mel Frequency Cepstral Coefficients (MFCCs), Spectral Centroid (SC), Spectral Roll-off Point (SR), Zero Crossing Rate (ZCR), and Spectral Kurtosis is considered as features.

2.3.1 Mel Frequency Cepstral Coefficients(MFCCs)

In the Mel Scale listeners usually respond to changes in pitch which seems to be equidistant from one another along this scale. Changes in frequency which corresponds to audible changes are easily reflected in this Mel Spectrogram when compared with typical frequency Spectrogram. The non-uniform spacing in frequency bands for Mel Spectrogram when compared with uniform and equal spacing bands of a normal Spectrogram yields a higher resolution and in the starting phase of cough sounds i.e., explorative phase where the glottis has just opened accounts for maximum energy where lower frequencies are considered. The normal frequency scale is converted into the Mel based frequency scale through the equation as given below:

$$fmel(f) = 2595 \times \left(1 + \frac{f}{700}\right) \quad \cdots\cdots\cdots (1)$$

The Mel spectrum obtained for audio cough samples undergoes Cepstral Analysis in order to compute the Cepstral coefficients, commonly referred to as Mel Frequency Cepstral Coefficients (MFCC)[12]. For every sample the feature extracted yields an M×N matrix where each row corresponds to MFCC features for distinct specific frame and each column a particular frame of the signal. Frame number *N* usually differs from one sample to another. Finally, to compute the MFCC coefficients triangular filter bank is followed by logarithmic compression and discrete cosine transform (DCT).The formula for computing the MFCCs is given in the equation 2.

$$C(n) = \sum_{m=1}^{M} \left[\log Y(m) \cos[\frac{\pi n}{M}(m - \frac{1}{2})\right] \quad \cdots\cdots\cdots (2)$$

2.3.2 Spectral Centroid (SC)

This feature is basically used to evaluate mean of the spectral energy which helps to find location of dominant formant frequency in each sub-band. The signal frequency changes and phase content over time can also be indicated through this feature.

2.3.3 Spectral Roll-off(SR)

Basically this feature is used to differentiate voiced sounds from unvoiced sounds. It describes slope of the signal's spectrum. 90th percentile of the power spectral distribution and skewness of signal's spectrum can also be obtained through this feature.

2.3.4 Zero-Crossing rate (ZCR)

The frequency with which an audio signal normally changes its sign from positive to negative and vice-versa within one frame is computed using this feature which indicates the variability in a signal [13]. This can be used to estimate dominant frequency component of the signal which is indicated in the equation mentioned in equation 3.

$$ZCR = \frac{1}{T-1}\sum_{t=1}^{T-1} \mu(s_t s_{t-1} < 0) \quad \cdots\cdots\cdots (3)$$

whereμ=1 when sign of $s_t and\ s_{t-1}$are different and μ=0 when $s_t and\ s_{t-1}$have same sign.

2.3.5 Kurtosis

This feature indicates the higher amplitudes occurrence for samples of audio signal in this study [14]. It also measures the tailedness of probability density which is given by the equation 4.

$$\Delta_x = \frac{E[(x_i[k]-\mu)^4]}{\sigma^4} \quad \cdots\cdots\cdots (4)$$

3. Classification Process

In the proposed work, different machine learning algorithms like Logistic Regression (LR), Support Vector Machines (SVM), Decision Tree and K-Nearest Neighbor (KNN) are used for classification and performance of each one of them is determined based upon their accuracy.

For all audio recordings, sampling rate is 44.1 KHz hence varying the frame lengths from $2^8\ to\ 2^{12}$ i.e., 256 to 4096, features can be extracted from frames with duration from 5 to 100 msec.Cough samples have different phases which carry essential features and has been segmented into chunks varying from 50 to 150 with step size of 20 to 30. The hyper parameters used in feature extraction process are shown in table 2.

TABLE 2 Hyper parameters used in feature extraction process

Hyper parameter	About	Range
MFCC number	Lower order MFCCs to be kept	$14 \times k, where\ k = 1,2,3,4,5$
Frame Size	Audio segmented into chunks called frames	$2^k where\ k = 8, \ldots, 12$
No of Segments	Grouping of frames into this number	$10 \times k, where\ k = 5,7,10,12,15$

3.1 Classifier Techniques

3.1.1 Logistic Regression (LR)

It is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0(stands for failure/no).

Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.

3.1.2 Support Vector Machines (SVM)

These are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression. But generally, they are used in classification problems. SVMs have their unique way of implementation as compared to other machine learning algorithms. Lately, they are extremely popular because of their ability to handle multiple continuous and categorical variables.

An SVM model is basically a representation of different classes in a hyperplane in multidimensional space. The hyperplane will be generated in an iterative manner by SVM so that the error can be minimized. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH).

3.1.3 K-Nearest Neighbor (KNN)

This algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems in industry. The following two properties would define KNN well −

KNN is a lazy learning algorithm because it does not have a specialized training phase and uses all the data for training while classification.

KNN is also a non-parametric learning algorithm because it doesn't assume anything about the underlying data.

K-nearest neighbours (KNN) algorithm uses 'feature similarity' to predict the values of new data points which further means that the new data point will be assigned a value based on how closely it matches the points in the training set.

3.1.4 Decision Tree

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

In decision analysis, a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated.

3.2 Cross Validation

The classifiers used in the proposed work have been trained and validated using K- fold cross validation. In the proposed work, 10 fold cross validation was implemented for avoiding the phenomenon of over fitting. This procedure has the provision for all patients in the used dataset which is small to be used for both training and testing purpose while ensuring there is strict maintenance of patient wise separation between cross validation folds.

3.3 Classifier Evaluation

The performance metrics of Sensitivity, Specificity, Precision, and area under curve (AUC), F-1 Score are used for evaluation of performance of the classifiers [15]. The normalized confusion matrix and performance metrics are evaluated so as to determine the best classifier model to be finally implemented for classification purpose.

4. Results and Discussions

Classification performance metrics for the collected cough samples is shown in Table 3. SVM model yields maximum accuracy of 96.9% in correctly classifying the cough samples for contaminated and healthy individuals, and with sensitivity of 96.7%, precision of 99.1%, F1 Score of 97.92%. Fig 12 also illustrates the fact through ROC curve analysis in which SVM model classifier has an AUC of 0.98.

Fig 8-11 demonstrates the confusion matrix plot for classifiers implemented in predicting an infected individual and healthy individual correctly referred to as 'True positive' and 'True Negative' cases in which SVM leads with an accuracy of 90.9% and 99.2%.

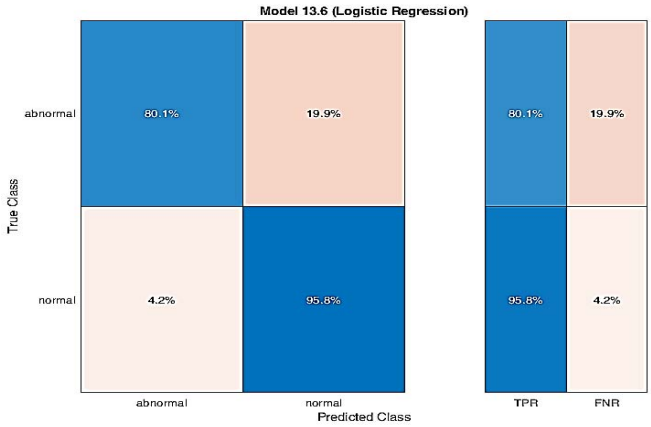

Fig 8. Confusion Matrix plot for Logistic Regression technique.

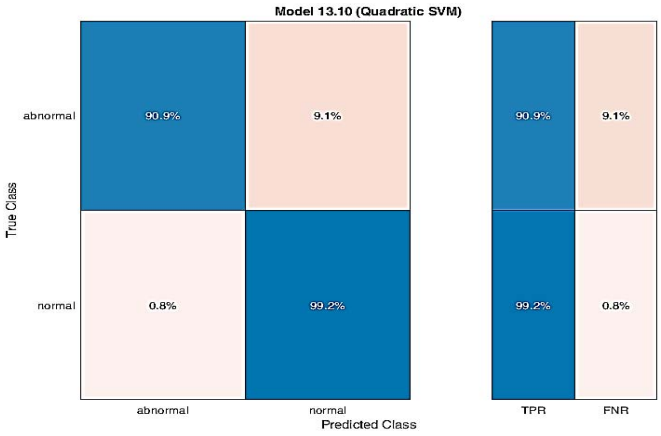

Fig 9. Confusion Matrix plot for SVM based classifier technique.

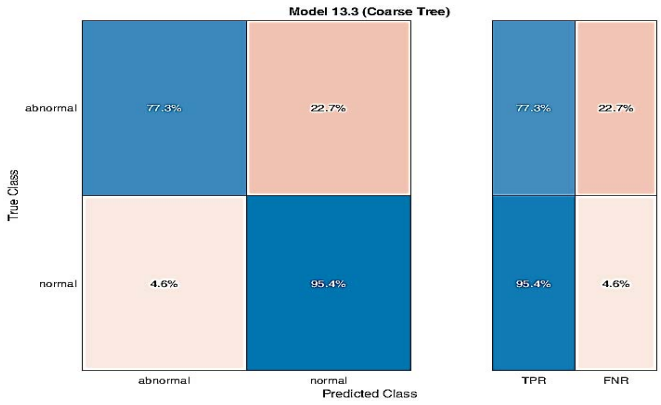

Fig 10. Confusion Matrix plot for Tree based classifier technique.

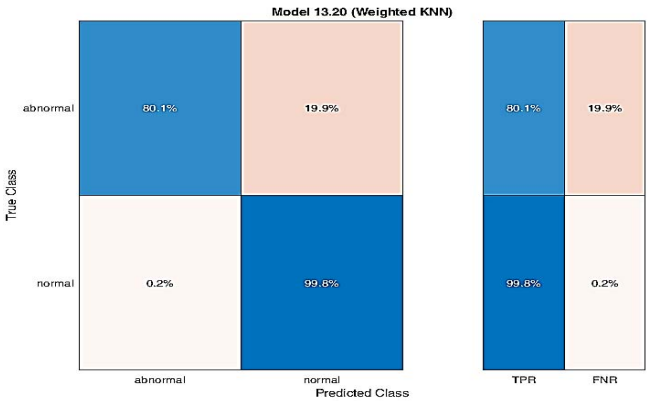

Fig 11. Confusion Matrix plot for KNN based classifier technique.

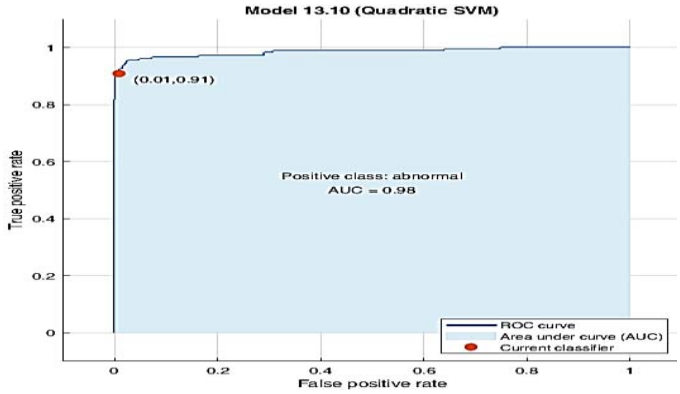

Fig 12. ROC curve for Quadratic SVM classifier with an AUC of 0.98

5. Conclusion and Future Work

In the current study COVID-19 cough based classification has been designed using audio samples collected from open search database and several machine learning architectures. The dataset comprised of subjects from 6 different continents except Africa. Once the preprocessing and feature extraction step was done, these were trained on 5 different classifier techniques and evaluated using the 10 fold cross validation technique. The proposed work shows that Quadratic SVM Model was best in terms of performance while differentiating COVID-19 cough from healthy ones with an AUC of 0.98.The proposed work with Machine learning requires features to be explicitly provided and select as to which feature results in building the classification model with maximum accuracy. The proposed approach of cough based classification towards non-Covid individuals can be extended with other respiratory illnesses as well like bronchitis, pertussis, asthma etc. so as to correctly distinguish coughs due to COVID-19 and these respiratory ailments. The more practical solution can be reached by using large dataset samples using neural network and deep learning algorithms.

.

TABLE 3 Performance Metrics for COVID_19 Diagnosis using several classifier techniques.

Classifier	Sensitivity (%)	Specificity (%)	Precision (%)	Area under Curve (%)	F-1 Score (%)	Accuracy (%)
KNN	93.16	99.29	99.79	98	96.36	94.5
SVM	96.73	97.56	99.16	98	97.92	96.9
LR	92.90	87.57	95.81	95	94.33	91.6
Decision Tree	91.93	86.07	95.39	87	93.62	90.5

REFERENCES

[1]. WHO et al., “Summary of probable sars cases with onset of illness from 1 November 2002 to 31 July 2003,” http://www.who.int/csr/sars/country/table2004 04 21/en/index.html, 2003.

[2]. D. Wang, B. Hu, C. Hu, F. Zhu, X. Liu, J. Zhang, B. Wang,H. Xiang, Z. Cheng, Y. Xiong et al., “Clinical characteristics of138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China,” JAMA, vol. 323, no. 11, pp. 1061–1069, 2020.

[3]. A. Carf`ı, R. Bernabei, F. Landi et al., “Persistent symptoms inpatients after acute COVID-19,” JAMA, vol. 324, no. 6, pp. 603–605, 2020.

[4].Post Washington. Hospitals are overwhelmed because of thecoronavirus.2020. https://www.washingtonpost.com/opinions/2020/03/15/hospitals-are-overwhelmed-because-coronavirus-heres-how-help/ Accessed on: Mar. 31, 2020. [Online].Available.

[5]. (2020, Nov.) COVID-19 dashboard by the centre for systems science and engineering (csse).John Hopkins

University.[Online].Available:https://coronavirus.jhu.edu/map.html

[6]. World Health Organization. 2020. ‘‘Report of the WHO-China joint mission on coronavirus disease 2019 (COVID-19) 2020. [Google Scholar].

[7]. National Institute of Health New coronavirus stable for hours on surfaces. 2020. https://www.nih.gov/news-events/news-releases/new-coronavirus-stable-hours-surfaces Accessed on: Mar. 31, 2020. [Online].Available.

[8]. P. Afshar, S. Heidarian, F. Naderkhani, A. Oikonomou,K. N. Plataniotis and A. Mohammadi, "COVID-CAPS:A Capsule Network-based Framework for Identification of COVID-19 cases from X-ray Images, "pp. 1-5, April 2020.

[9].L. Wang and A. Wong, "COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection ofCOVID-19 Cases from Chest X-Ray Images," pp. 1-12, May 2020.

[10]. G. Deshpande and B. Schuller, "An Overview on Audio, Signal, Speech, & Language Processing forCOVID-19," pp. 1-5, May 2020.

[11]. N. Sharma, P. Krishnan, R. Kumar, S. Ramoji, S. R. Chetupalli, P. K.Ghosh, S. Ganapathy et al., "Coswara–a database of breathing, cough, and voice sounds for COVID-19 diagnosis," arXiv preprintarXiv:2005.10548, 2020.

[12]. Wei Han, Cheong-Fat Chan, Chiu-Sing Choy, and Kong-Pang Pun,"An efficient MFCC extraction method in speech recognition," in IEEE International Symposium on Circuits and Systems, 2006.

[13]. R. Bachu, S. Kopparthi, B. Adapa, and B. D. Barkana, "Voiced/unvoiced decision for speech signals based on zero crossing rate and energy," in Advanced techniques in computing sciences and software engineering. Springer, 2010, pp. 279–282.

[14]. L. T. DeCarlo, "On the meaning and use of kurtosis." Psychological methods, vol. 2, no. 3, p. 292, 1997.

[15]. T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.