



# Machine learning approach for detecting Covid-19 from speech signal using Mel frequency magnitude coefficient

Sudhansu Sekhar Nayak<sup>1</sup> · Anand D. Darji<sup>1</sup> · Prashant K. Shah<sup>1</sup>

Received: 6 December 2022 / Revised: 9 February 2023 / Accepted: 19 February 2023 / Published online: 25 March 2023  
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2023

## Abstract

The Covid-19 pandemic is one of the most significant global health concerns that have emerged in this decade. Intelligent healthcare technology and techniques based on speech signal and artificial intelligence make it feasible to provide a faster and more efficient timely detection of Covid-19. The main objective of our study is to design speech signal-based noninvasive, low-cost, remote diagnosis of Covid-19. In this study, we have developed system to detect Covid-19 from speech signal using Mel frequency magnitude coefficients (MFMC) and machine learning techniques. In order to capture higher-order spectral features, the spectrum is divided into a larger number of subbands with narrower bandwidths as MFMC, which leads to better frequency resolution and less overall noise. As a consequence of an improvement in frequency resolution as well as a decrease in the quantity of noise that is included with the extraction of MFMC, the higher-order MFMCs are able to identify Covid-19 from speech signals with an increased level of accuracy. The procedures for machine learning are often less complicated than those for deep learning, and they may commonly be carried out on regular computers. However, deep learning systems need extensive computing power and data storage. Twelve, twenty-four, thirty, and forty spectral coefficients are obtained using MFMC in our study, and from these coefficients, performance is accessed using machine learning classifiers, such as random forests and *K*-nearest neighbor (KNN); however, KNN has performed better than the other model with having AUC score of 0.80.

**Keywords** Machine learning · Covid-19 · Mel frequency magnitude coefficient · Speech feature · Speech signal

## 1 Introduction

Speech is the most natural way for individuals to communicate with one another. The speech signal provides several layers of information, including what is spoken, the language being spoken, the speaker, their gender, and their feelings, among other things. Due to the fact that it is noninvasive, research on the automated evaluation of speech disorders based on pathological speech has garnered a growing amount of attention over the last decade. Speech pathologists have discovered a correlation between specific medical conditions

and the challenges experienced by individuals with their voices. The presence of pathological vocal disorders might potentially have a major effect on the individual's overall speech quality. Speech pathologists have shown that some medical conditions are associated with voice impairment. Cancer, asthma, Alzheimer's disease, Parkinson's disease, depression, schizophrenia, and autism spectrum disorder are just a few examples of the diseases that are included in this group. Researchers are increasingly interested in speech-based pathology, despite the fact that it is noninvasive. In recent years, there has been substantial development in the area of speech pathology, which has contributed significantly to the improvement of medical diagnosis.

When an individual's organs, such as the brain, heart, lungs, vocal folds, or muscles, are afflicted by illness, this might result in a change in the individual's voice. Consequently, speech analysis with artificial intelligence creates new prospects in the healthcare industry. The fields of speech and audio analysis both have numerous potential uses for AI. It might be included into the screening and early identifica-

✉ Sudhansu Sekhar Nayak  
d21ec005@eced.svnit.ac.in

Anand D. Darji  
add@eced.svnit.ac.in

Prashant K. Shah  
pks@eced.svnit.ac.in

<sup>1</sup> Sardar Vallabhbhai National Institute of Technology, Surat, Gujarat, India

tion of infected individuals, so contributing to the control and reduction in the number of infected individuals. In fact, voice technology is now largely recognized as one of the most promising areas, with healthcare predicted to become the leading user of voice applications within the industry. In more recent times, a substantial amount of research work has evolved with the goal of using respiratory sounds (such as coughs, breathing, and voice) as major sources of information in the context of the Covid-19 pandemic. Covid-19 is a respiratory disease that impairs breathing and voice, generating symptoms such as a dry cough, sore throat, an overly breathy voice, and usual breathing patterns [1].

The World Health Organization estimates that by the middle of 2022, almost 6 million people would have lost their lives as a consequence of the pandemic. The most common and effective test for Covid-19 is the reverse transcription-polymerase chain reaction (RT-PCR) test which has a number of drawbacks in addition to having a high degree of accuracy. A few of the drawbacks include the inconvenient nature of having to go to a specialized testing center, the necessity for knowledgeable medical personnel, the use of single-use testing kits, the intrusive nature of the operation, and the delayed wait for results [2]. Speech signal processing has recently been used in a variety of clinical aspects for the non-invasive diagnosis of problems, which has implications for both the effectiveness of remote health monitoring and the development of remote healthcare infrastructure. The capacity to record high-quality speech and other multimedia data, as well as analyze it locally or remotely, has made smartphones and other portable devices widely used in recent years which leads to a growing interest in automatically extracting information from speech for healthcare. Patients diagnosed with Covid-19 often have a vocal impairment, which serves as an early sign of the condition. The primary goal of such a noninvasive tool is to offer additional information to the clinical expert in order to reduce bias in the final diagnosis. Covid-19 is an infectious illness caused by a viral infection that can affect both the nose and the throat. Due to the fact that speech is produced by the vocal tract after the linear filtering of information from the source of excitation, the qualities of speech are impacted by the throat and nose while Covid-19 is taking place. The objective of the proposed research is to create a classification model that is based on machine learning and has the ability to reliably identify whether a person has a Covid or a non-Covid based only on their speech.

This paper is organized as follows: The literature review of Covid-19 detection using speech signal in Sect. 2. Section 3 explains the database we used. Section 4 goes through the proposed method for classifying speech as either Covid or Non-Covid which includes preprocessing, feature extraction and classification in detail, whereas Sect. 5 discusses the methodology and evaluation criteria. Section 6 gives results and discussion, and finally, Sect. 7 illustrates the conclusion.

## 2 Related work

In light of the time and money savings that may be made by avoiding medical testing, detecting Covid-19 via the analysis of speech signals has the potential to be a useful and cost-effective technique. A unique audio feature called C-19CC [3] is proposed and used to identify Covid-19, and the effectiveness of the method is assessed using two common speech datasets such as Coswara dataset and crowdsourced respiratory dataset. Aly et al. [4] proposed techniques for the sound-based diagnosis and detection of Covid-19 utilizing various respiratory sounds and human speech collected by cell phones. In order to accomplish this goal, deep learning models were used order to differentiate Covid-19 from healthy audio samples. Grant et.al [5] created a simple and efficient way that can be used to analyze sounds that may be used in a system to identify Covid-19 using a crowd-sourced database of audio recordings from self-identified Covid-19 positive and negative participants. Two alternative methodologies, random forest (RF) and deep neural networks (DNN), are used to analyze relative spectral-perceptual linear prediction (RASTA-PLP) and MFCCs features independently.

A dataset that has been pre-trained for three deep neural networks, such as CNN, LSTM, and Resnet50, includes noises such as coughing, sneezing, and speech, among other sounds and these deep neural networks serve as bottleneck feature extractors [6]. Despite the fact that all classifiers were capable of detecting Covid-19 coughs, the findings indicate that the ResNet50 classifier performed the best, with an area under the ROC curve (AUC) of 0.98 best differentiating Covid-19 positive coughs from healthy coughs. For the automated identification of Covid-19, Cai et al. [7] proposed an end-to-end network based on a transformer where features are extracted from the raw waveform and also created a feature extraction module where features are automatically extracted. To learn the sub-band decomposition of an audio signal, Dutta et al. [8] proposed a technique based on a parameterized convolutional neural network layer with a relevance weighting that uses cosine modulated Gaussian kernels.

Dash et al. [2] proposed a speech-based respiratory disease detection method for Covid-19 and Asthma, where various features, such as periodicity, spectral, cepstral, and spectral descriptors, are computed and then uniformly fused to obtain relevant statistical features, and these statistical features are then used as inputs for the gradient boosting algorithm.

## 3 Database

A crowdsourced database called Coswara [10] was created at the Indian Institute of Science (IISc) Bangalore. Recordings were made of each participant's coughing, breathing, and

voice sounds, information on their general health, gender, age, and a few pre-existing chronic diseases (such as diabetes and asthma), and their location. Quantification of the vocal variations brought on by the Covid-19 virus was performed in the preliminary study by continuously phonating the vowels A, E, and O, counting-fast and counting-normal.

We have utilized the dataset which is available until 24-02-2022, to train our model. The training data set consists of a total of 2089 random samples taken from each of the categories. All 1,594 of the people who were considered to be in good health were selected at random, and we assigned them to the ‘negative’ group. The recordings that were classified as ‘positive moderate,’ ‘positive asymptotic’ and ‘positive mild’ out of a total of 495 were classified as ‘positive,’ while ‘Res\_illness not identified,’ ‘Recovered full,’ ‘healthy’ and ‘No\_res\_illness\_exposed’ were classified as ‘negative.’

For the testing of our model, we make use of the same dataset as Coswara. The remaining 486 healthy persons were classified as being in a negative category. In addition, we chose the recordings that were classified as ‘positive moderate,’ ‘positive asymptotic’ and ‘positive mild’ out of the total of 168 samples and classified them as ‘positive.’ Based on the two distinct Covid-19 statuses, the samples are categorized, as indicated in Table 1.

## 4 Proposed method

Preprocessing the speech signal, feature extraction, and classification are the three stages of the proposed approach for analyzing and categorizing Covid and non-Covid from speech signals, as shown in Fig. 1

### 4.1 Preprocessing

We must do preprocessing on our recordings in order to increase the performance of our analysis. To do this, the recorded signal is separated into two parts: one containing the voiced region and the other including the silence or unvoiced regions [11]. Normalization, recognition of voiced regions, framing, and windowing are the steps involved in preprocessing. As a first step, the speech signal is normalized relative to its maximum possible value. Then, using a short-term energy approach, voiced sections are separated from the normalized speech signal. The voiced portions are divided into frames of 20msec length with a 50 percent overlap.

### 4.2 Feature extraction

Speech features are a collection of properties that are contained inside sound waves. The identification of these features is an essential stage that has an impact on the effectiveness of the system. The choice of appropriate features that

may represent the data in a way that is beneficial to the user is essential because it has the potential to have a substantial impact on the operation of the system.

#### 4.2.1 Mel frequency magnitude coefficient

In this proposed work, the feature extracted for the best parametric representation of speech signal is called Mel Frequency Magnitude Coefficients (MFMC) which is proposed by Ancilin and Milton [12]. The MFMC is obtained in the same way as the MFCC. However, there are two key differences: first, the magnitude of the fast Fourier transform (FFT) is used rather than the square of the magnitude. Second, the discrete cosine transform required in the MFCC extraction for decorrelation is not allowed. In order to get a greater resolution in the high-frequency components, the feature extraction process begins with pre-emphasis. The first thing that has to be done is to use a pre-emphasis filter on the signal so that the high frequencies may be amplified. The procedures that need to be carried out in order to obtain MFMC are depicted in Fig. 2.

It is possible to apply the pre-emphasis filter to the signal  $x(n)$  by making use of the first-order filter in the equation that is shown below.

$$y(n) = x(n) - a * x(n - 1) \quad (1)$$

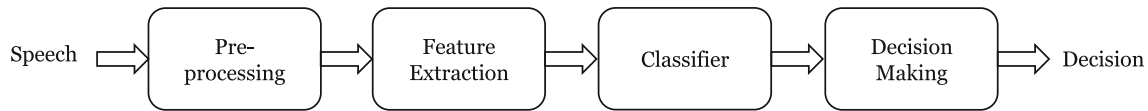
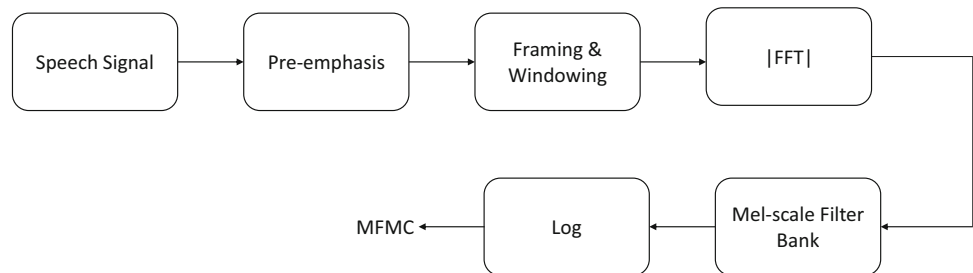
where  $y(n)$  is the output signal and the value of  $a$  is usually between 0.9 and 1.0. Applying Z transform to Eq. (1), then

$$H(z) = 1 - a * z^{-1} \quad (2)$$

The purpose of pre-emphasis is to make up for the high-frequency component of the sound that is muted as a result of the method through which people produce sound. The structure of the speech signal varies throughout time in response to the changes in the condition of the vocal tract. Therefore, in a broad sense, a speech signal is classified as a non-stationary signal [13]. As a result, the signal that has been pre-emphasized is filtered and then separated into short-time frames. If the frame duration is too large, the signal properties will shift inside the frame, thus care must be used while making the choice. A standard size for the duration of a frame is 20 milliseconds. A window is superimposed on the signal at each frame in order to smooth it out. In general, Hamming window is mostly used in speech signal processing to prevent the frames from discontinuities. The input speech signal is segmented into 20-ms frames by using a 50% window overlap criterion. A Hamming window must be applied to each frame before it can be multiplied in order to ensure the uniformity of the start and end positions at each frame. The equation that defines the output signal is  $y(n) = x(n) * w(n)$  where  $x(n)$  is input signal and  $w(n)$  is window.

**Table 1** Different Covid-19 cases in the Coswara dataset

Covid-19 status	Sample Category	No.of training samples	No.of testing samples
Positive	Positive_mild, Positive_asymp, Positive_moderate	495	168
Negative	Healthy, Resp_illness_not_identified, Recovered_full, No_resp_illness_exposed	1594	486

**Fig. 1** Proposed method for Covid and non-Covid speech classification**Fig. 2** MFMC feature extraction

The equation of Hamming window is given as

$$w(n) = 0.54 - 0.46 * \cos\left(\frac{2\pi n}{N-1}\right) \quad 0 \leq n \leq N-1 \quad (3)$$

Extraction of the MFMC feature is performed for each windowed frame of the speech signal.

It is possible to calculate the FFT of a windowed speech frame by applying Eq. (4).

$$X(k) = \sum_{n=0}^{N-1} e^{-j\frac{2\pi nk}{N-1}} \quad k = 0, 1, 2, \dots, N-1 \quad (4)$$

where  $N$  refers to the total number of samples included in the speech frame.

$N$  is dependent upon the sampling frequency as well as the length of the frame. The  $k$ -th Fourier transform coefficient is represented by  $X(k)$ , while  $n$ -th speech sample is represented by  $x(n)$ . Every Fourier transform coefficient, denoted as  $X(k)$ , is a complex number that possesses both phase and magnitude. In Eq. (5), the frame duration refers to the whole time, which includes any padded zeros that may have been used.

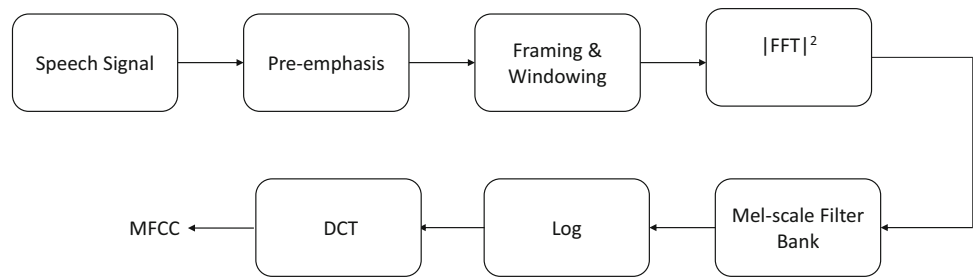
$$f = \frac{\text{Sampling rate}}{\text{Frame duration in samples}} * k \quad (5)$$

The next thing is that magnitudes of the Fourier transform coefficients  $X(k)$  is determined where  $k$  varies from

0,1,2,... $N-1$ . In MFCC, squared magnitude is utilized; this creates substantial variance if there are large signal components. The magnitude square will grow to a very large value for a sample  $X(k)$  that is much bigger than 1 when it is utilized, as it is in the case of MFCC, and this very high value will have a predominating influence on increasing the sum and the average values. If  $X(k)$  is smaller than 1, the square of  $X(k)$  will become extremely small, and its impact on the total and average values will decrease accordingly. Due to this substantial increase and a decrease in variety, the recognition rate is decreased. However, when the magnitude is considered, the influence of the fluctuation is shown to be almost insignificant. The magnitude spectrum is obtained from the Mel spectrum. This is presented using the Mel-frequency scale. Linear frequency converted to nonlinear Mel frequency through the use of an Eq. (6)

$$F = 1125 \ln\left(1 + \frac{f}{700}\right) \quad (6)$$

where  $f$  stands for normal frequency and  $F$  stands for Mel-frequency. The Mel-spectrum is then divided into  $M$  uniform bands that between the lower and upper frequency limits of zero and half of the sampling frequency, respectively, contain 50% of the Mel scale's overlapping region. Now, the  $M$  bands that are uniformly distributed over the Mel scale are transformed back into a linear frequency scale using Eq. (7). The linear frequency scale causes the Mel scale's uniform bands to become non-uniform. The bands that are found in the low frequency areas have a narrower bandwidth, while

**Fig. 3** MFCC feature extraction

the bands that are found in the high frequency area have a wider bandwidth. When it comes to frequency resolution, this non-uniform bandwidth provides better results in the lower frequency ranges than it does in the higher frequency ranges.

$$f = 700 \left( e^{\frac{m}{1125}} \right) - 1 \quad (7)$$

Consequently, triangular windows are utilized to modulate non-uniform bands. Equation (8) defines the triangle window for the  $m$ -th band of the  $M$  bands

$$H_m(k) = \begin{cases} \frac{k-f_a}{f_b-f_a} & f_a \leq k \leq f_s \\ \frac{f_c-k}{f_c-f_b} & f_b \leq k \leq f_s \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

After that, the total magnitudes of the frequency components of each triangular windowed band are summed together, and this step is represented mathematically by Eq. (9)

$$S(m) = \sum_{k=0}^{P_m-1} |X_m(k)| H_m(k) \quad 1 \leq k \leq m \quad (9)$$

where  $p_m$  denotes the number of frequency components present in the  $m$ -th band.

Logarithm is the mathematical operation that best fits the response of human ears to the decibel scale. Human hearing is measured on a scale based on decibels. By computing the logarithm of the sum of the Mel band magnitudes, the MFMC can be determined. For the  $m$ -th frequency band, the logarithm of the sum of the magnitudes of the individual frequency components is shown by Eq. (10)

$$MFMC(m) = \log_{10} \left( \sum_{k=0}^{P_m-1} |X_m(k)| H_m(k) \right) \quad 1 \leq k \leq m \quad (10)$$

#### 4.2.2 Mel frequency cepstral coefficient

The MFCC is considered to be the most important characteristic of all aspects of speech signal processing, including speech pathology and speech emotion detection. The MFCCs are extracted using the principles underlying human sound perception [14–17]. The procedures involved in obtaining the MFCC are explained in Fig. 3. The following number of steps are used to extract MFCC features from a windowed speech frame. (1) a spectrum is determined using a discrete Fourier transform, (2) power spectrum is computed, (3) Changes are made to the power spectrum's linear frequency scale so that it corresponds with the Mel frequency scale, (4) The Mel scale's frequency range is divided up into individual bands of equal width, (5) The frequency range provided by the Mel scale is divided into distinct bands that are all the same width, (6) The triangular window is applied to each band individually, and the energy of each band is determined by adding up its individual power components, (7) The energy levels in each band are logarithmically calculated, and (8) MFCC is obtained by calculating the DCT of the log energy.

#### 4.3 Classifiers

Several machine learning algorithms are used in order to identify the approach that is found to be the most effective in differentiating between a normal voice and a voice indicative of a pathological condition. In order to accomplish the classification task, we have used four different classifiers such as DT, RF, KNN, and SVM where each of which is developed with the help of the scikit-learn Python package[18] but KNN and RF outperformed the others. To overcome the imbalanced data, we have tried undersampling, oversampling, synthetic minority oversampling technique (SMOTE) and SMOTE-Tomek [19–21]. However, SMOTE outperforms the other data balancing techniques. Instead of duplicating existing minority class samples, this method seeks to create new ones.

#### 4.4 Random forest classifier

RF has quickly become one of the most widely used approaches in the industry due to its capacity to function



well on a large scale [9]. In RF, individual decision trees are built by combining bootstrap samples (sampling with replacement) with a randomly selected subset of attributes that do not include replacement. This is known as sampling without replacement. Adjusting the number of estimators or models that are used in the RF is one way in which the maximum capacity of the model may be enhanced.

#### 4.5 K-nearest neighbor (KNN) Classifier

A supervised machine learning approach called KNN is used to classification and regression issues [22,23]. By comparing the current observation to the training set, the KNN technique may determine which previous records are most similar. When categorizing a new observation, it is possible to take into account a set of neighbors known as  $K$ . This helps ensure that the classification is as accurate as possible.

### 5 Performance evaluation measure

The performance of classifier is measured using sensitivity (Sens), specificity (Spec) and receiver operating characteristics area under the curve (ROC-AUC) to measure their ability to distinguish between Covid patients and non-Covid participants. AUC is a widely used measure to evaluate the quality of models, especially in the context of binary classification issues [5,21,24]. Generally speaking, a higher AUC indicates that the model is more effective at distinguishing between the positive and negative classes.

### 6 Results and discussion

When categorizing pathological speech, it is standard procedure to first extract a different range of features, and then associate statistical functionals with each of those features. This creates utterance-level features, which may subsequently be used to categorize the speech. For the purpose of this study, we classified Covid and non-Covid speech signals using frame-level features. In this analysis, twelve, twenty-four, thirty, and forty spectral coefficients are considered from each voiced frame. The frame-level confidence score for both Covid and non-Covid categories is estimated using a machine learning techniques, such as RF and KNN. The confidence levels of each utterance are taken into consideration for the categorization of the number of frames, and the utterance was assigned to the group with the greatest sum of all confidence scores. When thinking about the frame-wise analysis of speech in order to extract features, we have taken into consideration the likelihood that certain sections of speech are more susceptible to such illnesses. When analyzing a healthy voice to a pathological one, many distinct

methods of machine learning were used in order to figure out which one delivers the best results.

When it comes to identifying the AUC score of each of the five distinct sounds, the MFMC feature performs much better than the MFCC feature which is shown in Table 2. It is shown that when utilizing two classifiers (RF and KNN), the AUC score of sound vowel-A by using the MFMC feature is 0.74 which is higher than the MFCC feature by considering spectral coefficients of order 12, 24, 30,40. Similarly, the AUC score of vowel-E and vowel-O, counting-fast, and counting-normal using the MFMC feature, are found to be 0.78, 0.80, 0.78, and 0.79 when the coefficients of order 12, 24, 30, 40 are considered, which is larger than the MFCC feature. In relation to the AUC score, which is presented in Table 2, the counting-fast phoneme class performs significantly more than all the other phoneme classes. When compared with the outcomes for other speech sounds, the classification scores for that counting-fast show several excellent outcomes.

Table 3 illustrates how well the proposed technique performs in comparison with other methods that are considered to be state-of-the-art when it is applied to the Coswara database. To put the best-performing results from speech in a wider context within the present research, we compare them to results from other methodologies. Deep learning-based algorithms were implemented by Aly et al. [4] in order to differentiate Covid-19 from healthy audio samples. The AUC score of the vowel-E was calculated to be 0.71, while the AUC score of the vowel-O was measured to be 0.75; both of these values are higher than the AUC of the vowel-A, which was calculated to be 0.61. Grant et al. [5] used two distinct features, MFCC and RASTA-PLP, which together yield the greatest performance of any feature set for speech. The average AUC is presented here for each of the fivefolds as a measurement of the overall performance of each classification technique

In this proposed work, Covid-19 identification from speech signals is carried out using the Mel frequency magnitude coefficient. MFMC varies from two characteristics of MFCC: The very first difference is that it uses magnitude rather than energy, and the other is that it does not use discrete cosine transform. According to the findings of the experiments, the MFMC performs better than the MFCC while classifying the detection of Covid-19. The DCT used in MFCC for de-correlating the log energy is not employed in MFMC. The spectrum pattern found in the log energy spectral coefficients is found using DCT in MFCC. Consequently, MFMC is a nonlinear log magnitude spectrum, while MFCC is a linear spectrum of a nonlinear log energy spectrum. During the process of extracting higher-order spectral characteristics such as MFMC, the spectrum is segmented into a larger number of subbands that have a smaller bandwidth. This narrowing of the bandwidth of the subbands results in improved frequency resolution as well as a reduced average

**Table 2** AUC, sensitivity, specificity score for MFCC and MFMC with different number of coefficients

Sounds	Coeffcs	MFCC						MFMC					
		RF			KNN			RF			KNN		
		Sens	Spec	AUC	Sens	Spec	AUC	Sens	Spec	AUC	Sens	Spec	AUC
VOWEL-A	12	0.54	0.63	0.62	0.54	0.63	0.65	0.52	0.65	0.63	0.51	0.64	0.71
	24	0.52	0.69	0.64	0.52	0.74	0.69	0.51	0.58	0.67	0.52	0.65	0.74
	30	0.50	0.70	0.62	0.51	0.70	0.67	0.50	0.61	0.64	0.53	0.67	0.68
	40	0.50	0.69	0.62	0.52	0.68	0.67	0.51	0.63	0.64	0.55	0.66	0.68
VOWEL-E	12	0.50	0.64	0.63	0.54	0.70	0.68	0.59	0.63	0.66	0.60	0.72	0.75
	24	0.51	0.72	0.64	0.51	0.71	0.69	0.57	0.67	0.67	0.54	0.69	0.74
	30	0.50	0.70	0.64	0.51	0.70	0.70	0.60	0.63	0.67	0.59	0.70	0.74
	40	0.50	0.69	0.63	0.50	0.71	0.69	0.57	0.66	0.67	0.54	0.65	0.73
VOWEL-O	12	0.58	0.64	0.65	0.54	0.70	0.67	0.63	0.65	0.67	0.63	0.71	0.72
	24	0.59	0.65	0.66	0.50	0.74	0.69	0.56	0.65	0.67	0.57	0.70	0.72
	30	0.58	0.64	0.66	0.50	0.75	0.69	0.57	0.65	0.67	0.55	0.69	0.71
	40	0.56	0.67	0.65	0.52	0.72	0.68	0.54	0.68	0.67	0.51	0.67	0.70
COUNTING-FAST	12	0.62	0.61	0.65	0.69	0.65	0.73	0.60	0.69	0.69	0.56	0.81	<b>0.78</b>
	24	0.61	0.59	0.65	0.68	0.66	0.73	0.58	0.67	0.68	0.70	0.71	<b>0.80</b>
	30	0.61	0.60	0.65	0.64	0.60	0.73	0.61	0.62	0.66	0.71	0.69	<b>0.78</b>
	40	0.62	0.59	0.65	0.67	0.65	0.72	0.59	0.60	0.66	0.79	0.62	0.79
COUNTING-NORMAL	12	0.58	0.65	0.64	0.60	0.74	0.71	0.64	0.59	0.65	0.55	0.81	0.76
	24	0.57	0.64	0.65	0.64	0.70	0.72	0.67	0.58	0.68	0.64	0.71	0.76
	30	0.60	0.62	0.64	0.61	0.67	0.71	0.65	0.58	0.65	0.69	0.67	0.76
	40	0.59	0.63	0.64	0.58	0.63	0.70	0.64	0.58	0.65	0.76	0.62	0.76

Bold number indicates the highest results among the all experimental results

**Table 3** Performance comparison of the proposed work with previous research papers

Research work	Feature	Classifier	AUC Score
Aly et al. [4]	MFCC	DNN	0.75
Grant et al. [5]	RASTA-PLP	RF	0.75
		DNN	0.71
		RF	0.69
		DNN	0.69
Proposed work	MFMC	KNN	<b>0.80</b>

Bold number indicates the highest results among the all experimental results

amount of noise. In order to increase the AUC score and shorten the extraction time, it is recommended to utilize the magnitude spectrum rather than the power spectrum and to directly use the log magnitude coefficients rather than the cosine-converted coefficients.

## 7 Conclusion and future work

In the field of healthcare, machine learning has been found to have the ability to have an effect, notably in finding and diagnosing diseases that are normally considered to be difficult to locate and diagnose. Because of this, a number of research are interested in developing methods that may diag-

nose the Covid-19 disease based on an analysis of speech. Our primary objective is to demonstrate the identification of Covid-19 by analyzing the speech signals. In this work, we compared the effectiveness of several ML approaches for identifying Covid-19 using speech analysis and calculated their AUC. To achieve this objective, we applied machine learning algorithms to differentiate Covid-19 from healthy audio recordings. For the purposes of model training, we extracted the feature, MFMC, from every audio sample that corresponds to distinct classes of sounds. After applying the various classifiers to the different audio signals, it was found that KNN performed better than the other classifiers in machine learning methods, which resulted in a score AUC of 0.80. The performance of MFMC would be further improved by using other categorization techniques, feature fusion, feature selection, and other approaches. Comparisons have been made between the suggested work and previous work of a similar kind, and it has been observed that the proposed technique yields more accurate results. To this end, we have been working to create a mobile health solution for the rapid detection of the Covid-19 condition, which might serve as a useful initial screening test and let patients track their symptoms before attending hospital visits. Using recordings of human speech gathered using mobile phones, this study diagnoses and detects the presence of Covid-19.

**Author Contributions** SSN conducted research and wrote the paper. ADD and PKS participated in the writing and preparation of the paper.

All authors reviewed the results and approved the final version of the manuscript.

**Funding** Not applicable.

**Data Availability** The data that support the findings of this study are openly available at <https://github.com/iiscleap/Coswara-Data>.

## Declarations

**Conflict of interest** The authors declare that they have no known competing financial interests or personal relationships.

**Ethical approval** Not applicable.

## References

1. Fagherazzi, G., Fischer, A., Ismael, M., Despotovic, V.: Voice for health: the use of vocal biomarkers from research to clinical practice. *Digit. Biomark.* **5**(1), 78–88 (2021)
2. Dash, T.K., Chakraborty, C., Mahapatra, S., Panda, G.: Gradient boosting machine and efficient combination of features for speech-based detection of covid-19. *IEEE J. Biomed. Health Inform.* **26**(11), 5364–5371 (2022)
3. Dash, T.K., Mishra, S., Panda, G., Satapathy, S.C.: Detection of covid-19 from speech signal using bio-inspired based cepstral features. *Pattern Recognit.* **117**, 107999 (2021)
4. Aly, M., Rahouma, K.H., Ramzy, S.M.: Pay attention to the speech: Covid-19 diagnosis using machine learning and crowdsourced respiratory and speech recordings. *Alex. Eng. J.* **61**(5), 3487–3500 (2022)
5. Grant, D., McLane, I., West, J.: Rapid and scalable covid-19 screening using speech, breath, and cough recordings. In: 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), pp. 1–6. IEEE (2021)
6. Pahar, M., Klopper, M., Warren, R., Niesler, T.: Covid-19 detection in cough, breath and speech using deep transfer learning and bottleneck features. *Comput. Biol. Med.* **141**, 105153 (2022)
7. Cai, C., Liu, B., Tao, J., Tian, Z., Lu, J., Wang, K.: End-to-end network based on transformer for automatic detection of covid-19. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) ICASSP 2022–2022, pp. 9082–9086. IEEE (2022)
8. Dutta, D., Bhattacharya, D., Ganapathy, S., Poorjam, A.H., Mittal, D., Singh, M.: Interpretable acoustic representation learning on breathing and speech signals for covid-19 detection. *arXiv preprint arXiv:2206.13365* (2022)
9. Södergren, I., Nodeh, M.P., Chhipa, P.C., Nikolaidou, K., Kovács, G.: Detecting covid-19 from audio recording of coughs using random forests and support vector machines. In: Interspeech 2021, Brno, Czechia, 30 August–3 September, 2021, International Speech Communication Association, pp. 916–920 (2021)
10. Sharma, N., Krishnan, P., Kumar, R., Ramoji, S., Chetupalli, S.R., Ghosh, P.K., Ganapathy, S., et al.: Coswara—a database of breathing, cough, and voice sounds for covid-19 diagnosis. *arXiv preprint arXiv:2005.10548* (2020)
11. Keerio, A., Mitra, B.K., Birch, P., Young, R., Chatwin, C.: On preprocessing of speech signals. *Int. J. Signal Process.* **5**(3), 216–222 (2009)
12. Ancilin, J., Milton, A.: Improved speech emotion recognition with mel frequency magnitude coefficient. *Applied Acoustics* **179**, 108046 (2021)
13. Rabiner, L.R.: Digital processing of speech signals, Pearson Education India, 1978
14. Sharma, S., Shukla, A., Mishra, P.: Speech and language recognition using mfcc and delta-mfcc. *Int. J. Eng. Trends Technol.* **12**(9), 449–452 (2014)
15. Warule, P., Mishra, S.P., Deb, S.: Classification of cold and non-cold speech using vowel-like region segments. In: 2022 IEEE International Conference on Signal Processing and Communications (SPCOM), pp. 1–5. IEEE (2022)
16. Deb, S., Warule, P., Nair, A., Sultan, H., Dash, R., Krajewski, J.: Detection of common cold from speech signals using deep neural network. *Circuits Syst. Signal Process.* **1**, 1–16 (2022)
17. Warule, P., Mishra, S.P., Deb, S.: Significance of voiced and unvoiced speech segments for the detection of common cold. *Signal Image Video Process.* **1**, 1–8 (2022)
18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)
19. Chamseddine, E., Mansouri, N., Soui, M., Abed, M.: Handling class imbalance in covid-19 chest x-ray images classification: using smote and weighted loss. *Appl. Soft Comput.* **129**, 109588 (2022)
20. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002)
21. Xia, T., Han, J., Qendro, L., Dang, T., Mascolo, C.: Uncertainty-aware covid-19 detection from imbalanced sound data. *arXiv preprint arXiv:2104.02005* (2021)
22. Chen, L., Wang, C., Chen, J., Xiang, Z., Hu, X.: Voice disorder identification by using hilbert-huang transform (hht) and k nearest neighbor (knn). *J. Voice* **35**(6), 932–e1 (2021)
23. Jayasanthi, M., Rajendran, G., Vidhyakar, R.: Independent component analysis with learning algorithm for electrocardiogram feature extraction and classification. *SIVIP* **15**(2), 391–399 (2021)
24. Pahar, M., Klopper, M., Warren, R., Niesler, T.: Covid-19 cough classification using machine learning and global smartphone recordings. *Comput. Biol. Med.* **135**, 104572 (2021)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.