Neural Computing and Applications (2021) 33:17621–17632
https://doi.org/10.1007/s00521-021-06346-3

ORIGINAL ARTICLE

Diagnosis of COVID-19 and non-COVID-19 patients by classifying only a single cough sound

Mesut Melek[1]

Received: 8 February 2021 / Accepted: 18 July 2021 / Published online: 30 July 2021

Abstract

In the last month of 2019, a new virus emerged in China, spreading rapidly and affecting the whole world. This virus, which is called corona, is the most contagious type of virus that humanity has ever encountered. The virus has caused a huge crisis worldwide as it leads to severe infections and eventually death in humans. On March 11, 2020, it was announced by the World Health Organization that a COVID-19 outbreak has occurred. Computer-aided digital technologies, which eliminate many problems and provide convenience in people's lives, did not leave humanity alone in this regard and rushed to provide a solution for this unfortunate event. One of the important aspects in which computer-aided digital technologies can be effective is the diagnosis of the disease. Reverse transcription-polymerase chain reaction (RT-PCR), which is a standard and precise technique for diagnosing the disease, is an expensive and time-consuming method. Moreover, its availability is not the same all over the world. For this reason, it can be very attractive and important to distinguish the COVID-19 disease from a cold or flu through a cough sound analysis via smartphones which have entered into the lives of many people in recent years. In this study, we proposed a machine learning-based system to distinguish patients with COVID-19 from non-COVID-19 patients by analyzing only a single cough sound. Two different data sets were used, one accessible for the public and the other available on request. After combining the data sets, the features were obtained from the cough sounds using the mel-frequency cepstral coefficients (MFCCs) method, and then, they were classified with seven different machine learning classifiers. To determine the optimum values of hyperparameters for MFCCs and classifiers, the leave-one-out cross-validation (LOO-CV) strategy was implemented. Based on the results, the k-nearest neighbors classifier based on the Euclidean distance (*k*NN Euclidean) with the accuracy rate, sensitivity of COVID-19, sensitivity of non-COVID-19, F-measure, and area under the ROC curve (AUC) of 0.9833, 1.0000, 0.9720, 0.9799, and 0.9860, respectively, is more successful than other classifiers. Finally, the best and most effective features were determined for each classifier using the sequential forward selection (SFS) method. According to the results, the proposed system is excellent compared with similar studies in the literature and can be easily used in smartphones and facilitate the diagnosis of COVID-19 patients. In addition, since the used data set includes reflex and unconscious coughs, the results showed that conscious or unconscious coughing has no effect on the diagnosis of COVID-19 patients based on the cough sound.

Keywords Cough sound · Classification · Machine learning · COVID-19 · Coronavirus · Computer-aided digital technologies

1 Introduction

A pandemic was declared by the World Health Organization on March 11, 2020. The cause of the disease was stated to be the new coronavirus 2 (SARS-CoV-2), which causes severe acute respiratory syndrome [1]. This epidemic disease, called COVID-19, affected the lifestyle, economy, social life, and education of billions of people. This disease, which is highly contagious and has no fully

This paper is dedicated to the memory of the late Dr. Mohsen MALEKI, who passed away from COVID-19 in November 2020.

✉ Mesut Melek
masoud.maleki1361@gmail.com;
mesutmelek@gumushane.edu.tr

[1] Department of Electronics and Automation, Gumushane University, 29100 Gumushane, Turkey

medically proven cure, has caused more than 1.94 M deaths worldwide by January 2021. The symptoms in patients with COVID-19 vary significantly depending on the individual, and it may take up to 14 days for the symptoms to appear [2]. Fever, fatigue, and dry cough are the most common symptoms [3] which can easily be mistaken for a cold or flu [2].

Since the day the pandemic started, healthcare teams around the world have been working on the diagnosis, follow-up, and treatment of patients. Moreover, most of the researchers are trying to help humans to go through these difficult days more easily by examining this event in their field [4–7]. In this case, studies on COVID-19 are expected from artificial intelligence (AI) and machine/deep learning (ML/DL) techniques. These techniques, which are generally known as computer-aided digital technologies, have affected and changed human life, especially in recent years. A system based on machine learning techniques is an intelligent system that gains experience from past occurrences and adapts to new situations without the need for explicit programming [8, 9]. These systems are used to solve a variety of computer science problems, from bioinformatics to image processing [10]. Therefore, machine learning systems and computer-aided digital technologies, in general, can be used on many fronts to combat COVID-19 [11, 12].

Early diagnosis of COVID-19 is important as many other diseases. The standard and definitive diagnosis of the COVID-19 is made via the reverse transcription-polymerase chain reaction (RT-PCR) test of the infected secretions from the nasal or throat cavity [13]. The results of this test can sometimes take up to 48 hours to come out. In addition, in order for the test to be effective, patients must remain isolated during this time. RT-PCR test is not only time-consuming but also expensive, and problems occur in large-scale deployments [14]. The cost of each test in the USA is approximately 23 dollars [15]. The governments try to test as much as they can every day, as they do not have the chance to test and control the whole country on one day. Moreover, the availability of the RT-PCR test is not the same all over the world. Therefore, it is of great importance to have a fast, simple, accurate, cheap, and easily accessible test.

One of the aspects in which machine learning can be effective is the early diagnosis of COVID-19. As mentioned previously, one of the most common and early symptoms of COVID-19 is coughing [2]. If non-coronavirus-induced coughs are distinguished from coronavirus-induced coughs through machine learning techniques, a cost-effective, easy, fast, and early diagnosis system can be offered. Since these systems can be installed on smartphones as an application or presented to users in a web-based environment, can be easily made available to everyone. In this case, in addition to the low cost, suspected candidates can record the cough sounds on their smartphones whenever they want and perform a preliminary determination for their status. Thus, great convenience will be provided in the early diagnosis of the disease. It can also reduce the burden of healthcare teams by effectively reducing the congestion in hospitals. This system, which is based on cough sounds, can also be used as a scanning method in airports, buses, waiting rooms of hospitals, nursing homes, and similar crowded environments [16].

In humans, different viruses, bacteria, or other acute and chronic health conditions, or even substances such as smoke and dust entering the lungs can cause coughing. In medicine, it is important for physicians to know whether the cough is wet, dry, or a wheezing, and whooping cough, in addition to how often and how severely the patient coughs [16]. Machine learning-based systems can detect the type of coughs (wet, dry, wheezing, and whooping cough) by providing medical professionals with more accurate clinical information about the frequency and severity of cough episodes. For this reason, even before the COVID-19 outbreak, studies on these issues were implemented.

Studies on cough sound can be divided into three categories. In the first group of studies, the aim is to distinguish cough sounds from other sounds. A preprocessing method was proposed for the detection of coughs in a noisy environment [17]. They used a fourth order Butterworth high pass filter in the pre-processing stage. Next, a methodology for automated analysis of cough sounds using support vector machines (SVM) was presented. In [18], a cough detection system that utilizes an acoustic onset detector in the preprocessing stage was proposed. This system, which is based on Long Short-Term Memory deep neural network architecture, has 90% sensitivity and 99% specificity. Barata et al. [19] conducted a study to ensure the scalability of existing cough detection models on various mobile devices. The authors investigated the performance of different methods across devices by recording 6737 cough samples and 8854 control sounds with 5 different recorders in a laboratory study with 43 subjects. Using an efficient convolutional neural network architecture and an ensemble-based classifier to reduce cross-device conflict, they achieved average accuracies between 85.9 and 90.9%. The proposed methods have demonstrated consistency across devices and the ubiquity of ubiquitous, scalable, and device-independent cough detection.

In the second group of studies, the aim is cough type classification (such as dry/wet). In a study [20], which was performed in 2011, two features that can be used to analyze cough sounds and distinguish between dry and wet cough sounds were identified. These features are the number of peaks of the energy envelope and the power ratio of two

frequency bands of the second phase of the cough signal. However, a clear distinction was observed by using only eight dry and eight wet cough sounds. In recent study [21], an objective approach based on the acoustic features of the cough sound collected by smartphones from 131 subjects was proposed. For classification between wet and dry coughs, sensitivity and specificity of the system were calculated as 88% and 86%, respectively. They obtained these values by classifying the features that they extracted from the time and frequency domain with a random forest classifier.

The third group of studies are the most comprehensive studies in this field. In these studies, cough sounds are first detected from other sounds and then classified into different types of coughs by the system. Pramono et al. [22] presented a system for diagnosing whooping cough in young children, which can be fatal if untreated. In their study, audio recordings from 38 patients were used for automatic diagnosis of pertussis by analyzing the cough and whooping sounds. The algorithm was able to successfully detect whooping cough from sound recordings and automatically detect individual cough sounds with 92% accuracy by using a logistic regression model. In [23], deep neural networks were used for cough detection based on the convolutional neural network and the recurrent neural network. The accuracy rate of the system was reported to be 82.5% for the three classes defined, namely, cough, speech, and other.

Since the start of the pandemic, researchers working on computer-aided digital technologies have offered different ideas, solutions, and methods based on previous experiences. This covers a range from the analysis of the CT scans and X-ray images [24–28] for the diagnosis of COVID-19 to emotional and sentiment analysis from social media [29–31]. Among these studies, it appears in studies based on sound analysis and especially cough sound. One of the parameters that has the greatest impact on machine learning studies is the data set. Since COVID-19 is a newly emerging disease and more importantly considering the status of the COVID-19 patients, it is very difficult to collect and access data sets. However, despite all these difficulties, studies on sound and especially on cough sound appear in the literature. Although most of these studies are not peer-reviewed yet, they can be obtained from different preprint banks. In [32], data obtained from lung auscultation with digital stethoscope were used for the diagnosis of covid-19. In [33], changes in the vocal patterns of Covid-19 patients were analyzed. In a similar study [34], audio recordings of COVID-19 patients were used to express the severity of the disease. Alsabek et al. [35] examined the speech signals of COVID-19 and non-COVID-19 patients by calculating the Mel-frequency cepstral coefficients (MFCCs). They used Pearson's correlation coefficients to show the relationship between the two signals.

In the literature, it is encountered in studies on the analysis of the cough sound. In [36], respiratory sounds of COVID-19 patients, with the help of a binary classifier, were distinguished from respiratory sounds of healthy people with an area under the curve (AUC) exceeding 0.80. The authors used the support vector machine (SVM) classifier in the classification stage. The used data set was gathered using a web-based app and an Android app. Ali et al. [37] presented a mobile application that records and analyzes 3-second cough sounds through an application called AI4COVID-19. A total of 328 cough sounds of four different types including COVID-19, asthma, bronchitis, and healthy from 150 people were recorded and classified. The authors used the MFCCs method in the feature extraction stage, and the accuracy rate of the system was calculated as 92.85%. In a study [38] on Coswara and Virufy data sets, features extracted from the frequency and time domain were classified using machine learning methods. The results were compared with the recurrent neural network (RNN) method, which is common in deep learning methods. The system's accuracy rate was calculated as 81.25% for RNN when two different databases were used for the training and testing set. In [39], the authors classified 81 cough sound recordings (8 COVID-19, 28 pneumonia, 15 pertussis, and 30 healthy) through machine learning and deep learning. In this four-class study, an overall accuracy rate of 94% was given for the SVM classifier. A total of 328 cough sounds collected from 200 patients at a hospital in New Delhi, India, were classified into four classes (COVID-19, Asthma, Bronchitis, and Healthy) via the Deep neural network (DNN) [40]. In this data set, in addition to cough sounds of 100 COVID-19 patients, breathing sounds, counting from 1 to 10, sustained phonation of 'a', 'e', 'o' sounds, demographic, fever, headache, sore throat were given. By using only cough sounds, in binary classification (COVID-19 vs. non-COVID-19), the system's accuracy rate, sensitivity, and specificity were calculated as 90.8%, 90.1%, and 90.3%, respectively. When cough data and symptoms data are used together, the accuracy rate of the system was computed as 96.5%. In [41], cough sounds that were collected from 3621 people via mobile phones were classified with 0.72 AUC by using deep convolutional neural networks (CNNs). In [42], cough sounds were classified with a 95.86% accuracy rate with SVM's RBF kernel function classifier by obtaining features by the MFCCs method. The sensitivity of the system to COVID-19 cough sounds was calculated as 98.6%, and the sensitivity was obtained as 91.7%.

In summary, classifying Covid-19 patients from non-Covid-19 patients by using cough sounds is a new area of

research. As it is said, when the literature is reviewed, we come across a small number of studies in this area. This shows that we are just at the beginning of this road. Most of the available works have either been submitted for conferences or have not yet been peer-reviewed. However, the main purpose of these studies, as stated in most of them, is to show that it is possible to distinguish COVID-19 patients from non-COVID-19 patients through cough sound analysis. These studies show that it is possible and meaningful to analyze the data recorded by different phones and microphones and in different environments on the same system. In our proposed study, it shows that the analysis of cough sounds is useful in the non-contact detection of COVID-19 and it is possible to distinguish COVID-19 and non-COVID-19 patients. Deep learning models were used in many of the mentioned studies. However, datasets containing Covid-19 coughs are small, and deep learning models for such small datasets are often overfitted [36]. Therefore, to avoid this problem, common machine learning methods were used in the proposed study. The features were obtained by the MFCCs method on the data set acquired by combining these two data sets and classified with seven different classifiers. The optimum values of the hyperparameters of the system were determined based on the LOO-CV strategy. Finally, effective features were determined separately for each classifier using the sequential forward selection (SFS) method.

Also, most of the studies conducted have been based on recordings involving a few coughs. That is, for example, a 9-second recording has 3 or 4 coughs. In addition, all of the studies have been conducted on mandatory and conscious cough sounds. In this study, considering these two important points, two different datasets were combined and used. The virufy [43] data set is a public data set and contains 121 single cough records. The novel coronavirus cough database (NoCoCoDa) [16] is available to researchers free of charge upon request. This data set contains 73 single coughs that include reflex COVID-19 cough sounds and are not mandatory. In this way, mandatory and conscious single cough sounds, in addition to reflex single cough sounds of COVID-19 patients, were successfully distinguished from single cough sounds of non-COVID-19 patients in the present study. The results showed that the proposed system is more successful than other systems. In addition, the results revealed that conscious or unconscious coughs have no effect on the diagnosis of COVID-19 patients with cough sounds. In other words, our study proved that the diagnosis of Covid-19 patients can occur without the need for unconscious coughs.

In the following section, materials and methods are explained. In the third section, the results are given and in the fourth section, the results are discussed. Section 5 presents the conclusion.

2 Materials and methods

2.1 Data set description

2.1.1 Virufy COVID-19 open cough data set

The virufy COVID-19 open cough data set is the first free, publicly available data set containing COVID-19 cough sounds [43]. COVID-19 PCR test results were also given along with the demographics of all the patients in the data set. After obtaining informed patient consent, the data were collected from patients in a hospital and under surveillance and verified by physicians, following standard operating procedures (SOPs). Some patients have no symptoms, while others have symptoms such as fever, chills, or sweating, shortness of breath, new or worsening cough, sore throat, loss of taste, loss of smell. These cough sounds, which were collected from 16 patients, were recorded at a sampling frequency of 48 kHz. Then, each recording was split so that it contained only one cough with the duration of 1.645 seconds. Thus, the data set consists of 121 single cough records, 48 of which were reported to have a positive PCR test result, and 73 were reported to have a negative test result. The original format of the records (before splitting) is also given in the data set.

2.1.2 NoCoCoDa

In [16], public media interviews with COVID-19 patients were manually reviewed and the cough sounds were separated one by one and recorded. These interviews were broadcasted online by news sources. This database, called the NoCoCoDa, contains a total of 73 single cough sounds and is available to researchers free of charge upon request. This data set, with a total of 13 interviews attended by 10 people, includes reflex COVID-19 cough sounds. The cough sounds were recorded as a .WAV file with a sampling frequency of 44.1 kHz. In addition to the data, information about the patients is given in an additional file. Since NoCoCoDa is derived from reports and news programs, other sounds such as speech or music are heard in the background in some cough recordings. In a few, a mixture of throat clearing and coughing was also found. All this is specified in the additional file.

In the present study, these noisy and suspicious coughs were removed from the NoCoCoDa data set and the remaining 59 coughs were used. After combining these two data sets, the distribution of cough sounds between the two classes is given in Table 1. As can be seen, a total of 180 cough sounds from 107 COVID-19 patients to 73 cough sounds from non-COVID-19 patients were used in this study.

2.2 Mel-frequency cepstral coefficients (MFCCs)

MFCCs are one of the popular and successful methods for obtaining features in voice analysis and automatic speech recognition systems [44]. MFCCs is a digital technical analysis that simulates the perception of human ears and is calculated on the basis of Fast Fourier Transform (FFT). Since the characteristics of speech signals remain stable in a very small time interval (about 20–30 ms), they are processed in very short time intervals [45, 46]. This short interval is called the frame. Frames are usually chosen to overlap to make transitions between frames smoother. Similar to the calculation of spectrogram, here, the windowing process takes place to avoid a discontinuity at the beginning and end of the frames. The commonly used window structure is Hamming. After windowing, FFT is applied to transform each frame from the time domain to the frequency domain. The mel unit is a unit designed to imitate the perceptual feature of the human ear. Conversion between the mel scale and the frequency scale is provided by the equation given below.

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

In this way, MFCCs are the expression of the short-time power spectrum of the sound signal on the mel scale [47, 48]. When MFCCs are calculated for a cough sound, a matrix is obtained in the M × N matrix, where M is the number of MFCCs and N is the number of segments (the number of frames).

In the literature, MFCCs was used for the classification of cough sounds. For example, in [20], the features were extracted by the MFCCs method for the classification of dry and wet coughs. In order to obtain features with the MFCCs method, attention should be paid to important factors, called hyperparameters, which include the type of window used, frame length, frame overlap length, number of segments used for feature extraction, and number of MFCCs. In this study, the chosen window type was Hamming, and the frame overlap length was half of the frame length. The optimum values of the other three hyperparameters (the frame length, number of MFCCs, and number of segments used for feature extraction) were chosen using the LOO-CV strategy.

Table 1 Distribution of cough sounds between the two classes

Data set	COVID-19	Non-COVID-19	Total
Virufy	48	73	107
NoCoCoDa	59	0	59
Total	107	73	180

2.3 Classification

Today, classification is used in various fields, from medical or genomic predictions to systems such as spam detection and face recognition, and even in finance [49]. In the classification process, a classifier is trained with samples with certain labels and a model is created. Then, the model is used to guess the label of unknown samples [50]. Many classifiers based on machine/deep learning methods were used in the classification of cough sounds. For example, in [51], logistic regression (LR), support vector machines (SVM), multilayer perceptrons (MLP), convolutional neural networks (CNN), long short-term memory (LSTM), and residual-based neural network architecture (Resnet50) were used.

In this study, popular classifiers in machine learning systems were used to classify COVID-19 and non-COVID-19 cough sounds. These are SVM, linear discriminant analysis (LDA), k-nearest neighbors (kNN), and partial least squares regression (PLSR). In SVM, LDA, and kNN classifiers, two different structures of the model were implemented. In SVM, two different nonlinear kernels, namely radial basis function and polynomial kernels, were used. In kNN, Euclidean and Chebyshev distance metrics, and in LDA, linear and quadratic decision surfaces were tested. By adding PLSR to these six classifiers, a total of seven different classifiers were created and the results of each classifier were calculated. To determine the values of the hyperparameters in each classifier, the LOO-CV strategy was used.

2.4 Measuring the performance of the system

The performance of a classification system can be measured with different metrics. In this study, the accuracy rate, AUC, F-measure, sensitivity, and specificity were used to measure the performance of the system. There are different strategies for calculating these metrics. One of the popular strategies is LOO-CV [52]. The LOO-CV strategy is adopted in a system when the number of samples in the data set or even just the number of samples in a class is low [53]. In this strategy, the data set containing N samples is divided into two sections. The N-1 sample is used for training the classifier, and the single remaining sample is used for testing the model. All the samples are used for testing only once, so the process is repeated N times and, in this way, different metrics can be computed. In this study, the LOO-CV strategy was used to calculate the metrics, taking into account the total number of samples (180 samples) in the two classes.

3 Results

A method was proposed based on machine learning to diagnose COVID-19 patients from non-COVID-19 patients by cough sounds. The proposed method was tested on a data set that includes virufy and NoCoCoDa data sets. The features were extracted from cough sounds using the MFCCs method and classified with seven different classifiers. To select values of hyperparameters in feature extraction and classification processes, the accuracy rate metric was calculated according to the LOO-CV strategy. In searching for the optimum value of a hyperparameter, all the other hyperparameters were kept constant. The value reaching the highest accuracy rate in the searched range was selected as the optimum value of that hyperparameter. In all of the steps of the study, for the MFCCs method, the window type was chosen as Hamming, and the frame overlap length was half of the frame length.

3.1 Determination of the values of hyperparameters in the feature extraction phase

As mentioned earlier, in the MFCCs method, three hyperparameters were taken into account. These are the frame length, number of MFCCs, and number of segments used for feature extraction. Frames lengths of 512, 1024, 2048, and 4096 samples were tested to determine the optimum frame length. In this case, the number of MFCCs was selected as 13, and the number of segments used for feature extraction was selected as N. That is, for each cough, a 13xN matrix was obtained, and by averaging in all N segments, a 13x1 feature vector was obtained. Then, the feature vectors were transferred to the classifiers and classified. The accuracy rate obtained for each frame length based on the LOO-CV strategy is separately presented for each classifier in Table 2. The hyperparameters of the selected classifiers also appear in the table. As can be seen, the Chebychev-kNN classifier successfully classified the cough sounds recorded from COVID-19 and non-COVID-19 patients for the 2048 frame length with an accuracy rate of 0.9389, followed by the Euclidean-kNN classifier with an accuracy of 0.9167. By looking at the results in general, all the classifiers achieved higher accuracy for the 2048 frame lengths. Therefore, for the continuation of the study, the optimum value of the frame length hyperparameter was chosen as 2048 samples.

To determine the optimum number of MFCCs, a scanning between 2 and 39 was performed. For this, as in the previous step, the number of segments used for feature extraction was selected as N. In this way, for example, when the number of MFCCs is 2, two features, and when it is 39, 39 features are extracted. The performance of all the classifiers was measured by the accuracy rate metric using the LOO-CV strategy. The classifiers' hyperparameter was adjusted as in the previous step. The results are given in Fig. 1. As it turns out, Euclidean-kNN achieved the best performance using 19 MFCCs with an accuracy rate of 0.9500 followed by Chebychev-kNN and polynomial-SVM with an accuracy rate of 0.9389. These ratios were obtained using 13 and 17 MFCCs, respectively. Therefore, the number of MFCCs was determined as 19 for the continuation of the study.

The last hyperparameter in the feature extraction phase is the number of segments used for this phase. In order to obtain the optimum segment number, numbers from 1 to 50 were used and the averages were obtained. For this purpose, only the first segment of the first 19-MFCCs was used as the feature vector and classified by classifiers. Then, 19 features obtained by the average of the first two segments of 19-MFCCs were classified. This process continued until 50. The hyperparameters of the classifiers were chosen as in the previous steps. The results are shown in Fig. 2. Euclidean-kNN appears to be the most successful classifier in using 17 segments, with an accuracy rate of 0.9833. After determining the optimum values of the hyperparameters in the feature extraction phase, the striking point is the approximately 7% increase in the accuracy of the Euclidean-kNN classifier.

3.2 Determination of the values of hyperparameters in the classification phase

To determine the optimum hyperparameter values of the classifiers, the accuracy rate metric obtained by the LOO-CV strategy was used. In the RBF-SVM classifier, the range of 0 to 3 was screened by steps 0.1 to determine sigma. The results are given in Fig. 3. When sigma=1.3, the accuracy rates reached 0.9611. In the polynomial-SVM classifier, the order hyperparameter was searched between 1 and 4. The highest accuracy rate (0.9556) was achieved when the third-order kernel function was used.

Different gammas from 0 to 1 were tested by steps 0.1 to determine gamma in the linear-LDA classifier. The results showed that the highest accuracy was at gamma = 0.6, as given in Fig. 4. The gamma was changed to 0 and 1 in the quadratic-LDA classifier, and the accuracy rate was calculated. A higher accuracy rate (0.9056) was calculated at gamma = 0.

In order to determine k in both types of kNN classifiers, the classification accuracy rate was calculated from 1 to 25 by steps 1. The results are given in Fig. 5. The highest accuracy rates at $k = 1$ were calculated for both classifiers. Thus, the accuracy rates did not change for these classifiers.

Table 2 Classification results for different frame lengths

Frame lengths (samples)	Polynomial-SVM Order=2	RBF-SVM Sigma =1	Linear-LDA Gamma=0	Quadratic-LDA Gamma=0	Chebychev-kNN K=1	Euclidean -kNN K=1	PLSR component=13
512	0.8667	0.8944	0.8944	0.8278	0.9056	0.9111	0.8278
1024	0.8722	0.8944	0.9000	0.8389	0.9222	0.9111	0.8444
2048	0.8722	0.8944	0.9056	0.8556	**0.9389**	0.9167	0.8556
4096	0.8667	0.8944	0.9000	0.8500	0.9056	0.9167	0.8500

The highest value is indicated in boldface

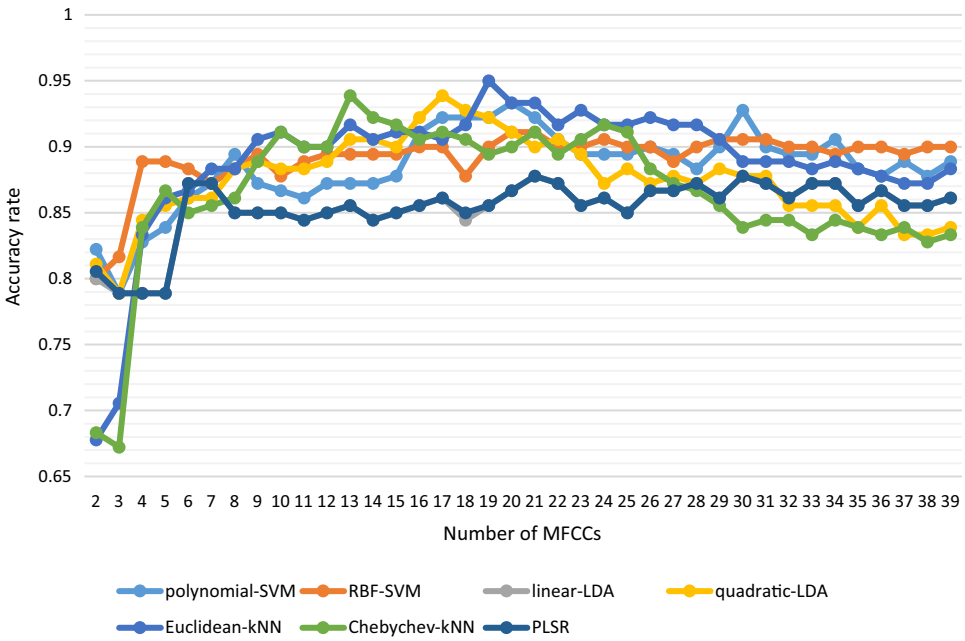

Fig. 1 Classification results for different numbers of MFCCs

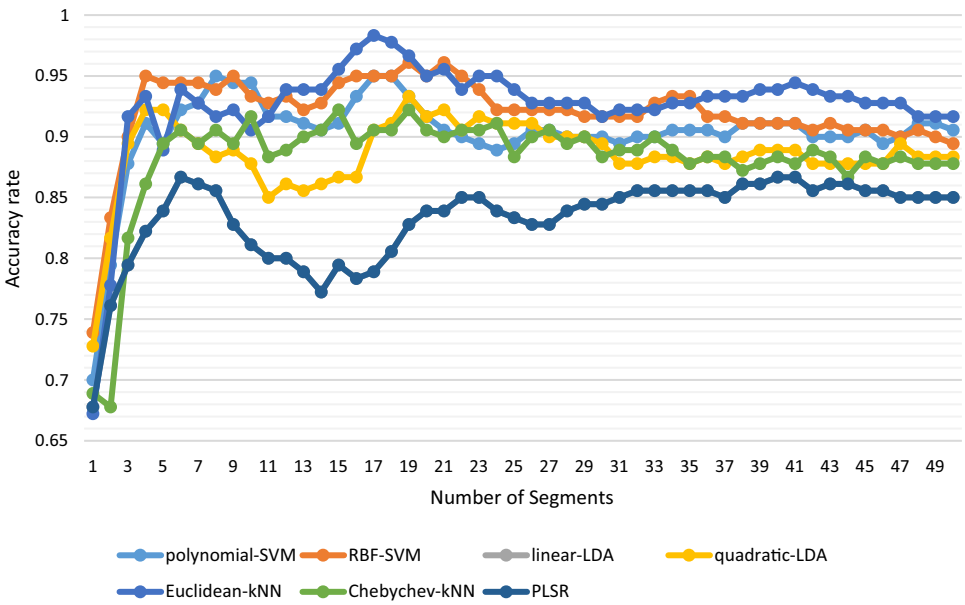

Fig. 2 Classification results for different numbers of segments used for feature extraction

Table 3 The results of classification by tuning hyperparameters in classifiers

Classifier	Hyperparameter	ACC	Sen. Non-COVID-19	Sen. COVID-19	*F*-measure	AUC
Polynomial-SVM	Order=3	0.9556	0.9452	0.9626	0.9452	0.9539
RBF-SVM	Sigma=1.3	0.9611	0.9583	0.9630	0.9517	0.9606
Linear-LDA	Gamma=0.6	0.8111	0.7808	0.8318	0.7703	0.8063
Quadratic-LDA	Gamma=0	0.9056	0.8082	0.9720	0.8741	0.8901
Euclidean-kNN	*K*=1	**0.9833**	**1.0000**	**0.9720**	**0.9799**	**0.9860**
Chebychev-kNN	*K*=1	0.9056	0.8904	0.9159	0.8844	0.9031
PLSR	Component =4	0.8111	0.7808	0.8318	0.7703	0.8063

The highest value in each metric is indicated in boldface

Table 4 Effect of the SFS method on classification results

Classifier	Used Features	ACC (after used SFS)	ACC (before used SFS)
Polynomial-SVM	19	**0.9556**	**0.9556**
RBF-SVM	18	**0.9667**	0.9611
Linear-LDA	11	**0.8388**	0.8111
Quadratic-LDA	15	**0.9111**	0.9056
Euclidean-kNN	19	**0.9833**	**0.9833**
Chebychev-kNN	13	**0.9444**	0.9056
PLSR	17	**0.8277**	0.8111

The highest value in each classifier is indicated in boldface

Table 5 Comparison of the results of the proposed study with the results of other studies

Study \ metrics	ACC	Sen. Non-COVID-19	Sen. COVID-19	*F*-measure	AUC
[40]	0.9080	0.9010	0.9030	0.9060	—
[51]	—	—	—	—	0.7200
[36]	–	–	–	–	0.8200
[38]	0.8125	–	–	–	0.8000
[37]	0.9285	0.9114	0.9457	0.9297	—
[42]	0.9586	0.9863	0.9167	—	—
[51]	0.9501	0.9800	0.9300	—	0.9632
Proposed method	**0.9833**	**1.0000**	**0.9720**	**0.9799**	**0.9860**

The highest value in each metric is indicated in boldface

Different components were tested by steps 1 from 2 to 19 in order to determine the component in the PLSR classifier. The results show that the highest accuracy rate was when the component = 4, as given in Fig. 6. Thus, the PLSR classifier classified the cough sounds of COVID-19 and non-COVID-19 patients with an accuracy rate of 0.8111, such as the linear-LDA.

In order to see the system performance more clearly, after determining the optimum hyperparameter values of the classifiers, in addition to the accuracy rate, four more metrics were calculated. The determined hyperparameter values and calculated metrics are given in Table 3. As it turns out, Euclidean-kNN is more successful than the other classifiers in all the metrics. The Euclidean-kNN classifier showed 0.9720 and 1.0000 sensitivity to the COVID-19 and non-COVID-19 class, respectively.

3.3 Feature selection based on SFS

In the last step of the study, the feature selection process based on the SFS method was performed separately for each classifier. The results of this step are given in Table 4. In order to see whether the SFS method has any effect, the accuracy rates calculated in the previous step are also

shown in the table. As seen, there appears to be an increase for all the other classifiers, with the exception of the Polynomial-SVM and Euclidean-kNN classifiers.

4 Discussion

In this study, for the first time, the cough sounds of conscious and unconscious COVID-19 patients (in the same class) were classified against the cough sounds of non-COVID-19 patients. The results showed that this process was successful by the Euclidean-kNN classifier with an accuracy rate of 0.9833. Thus, it was observed that conscious or unconscious cough sounds did not have any significance in the diagnosis of COVID-19. Because if there were a difference between conscious coughs and unconscious coughs, this difference would have been reflected in the frequency components of the cough sounds and would have an effect at the classification stage, thus causing a decrease in the classification results. So, our study proved that the diagnosis of Covid-19 patients can occur without the need for unconscious coughs.

As said earlier, few studies have been conducted to distinguish between COVID-19 and non-COVID-19 patients based on cough sounds in the literature. Most of the available works have either been submitted for conferences or have not yet been peer-reviewed. In Table 5, the results of these studies and the results of the proposed study are presented for comparison. Except for [36] and [42], other studies used deep learning methods for classification. Datasets containing Covid-19 coughs (as this is a new area of research) are small, and deep learning models for such small datasets are often overfitted [36]. Although popular machine learning methods were used in this study, as can be seen, all the metrics of the proposed study were higher than those of the other studies. An important issue to note is the sensitivity of the system in the diagnosis of COVID-19 patients. The proposed study is more successful than other studies with a sensitivity of 0.9720 to COVID-19 patients.

The proposed system can be installed on smartphones as an application or presented to users in a web-based environment. While presenting health technology app to the application stores, important issues such as application health, data security, and user safety should be considered [54]. Both Apple and Google have specific requirements that must be met when publishing apps. Stricter guidelines apply when it comes to healthcare practices, but there are no serious limitations or barriers in the use and publishing of these systems. Google Play generally provides a faster response and simpler guidelines in the approval process compared to the App Store [55]. While Android health app requirements are limited compared to other apps, it is your responsibility to comply with health regulations such as Health Insurance Portability and Accountability Act (HIPAA) compliance. More information on this topic can be obtained from [54]55.

Another important issue to be considered is the number of samples in the data set. The more samples used in machine learning-based systems, the more reliable the system will be, and generally, the larger the dataset, the greater statistical power for pattern recognition [56]. Since COVID-19 is a newly emerging disease, it is very difficult to collect and access data sets. However, there is no limit on this issue. There are only different validation strategies. In our study, we used the leave-one-out cross-validation strategy, which evaluates all possibilities.

5 Conclusion

To diagnosis the COVID-19 disease, it is essential to have a cost-effective, fast, easy, and accurate method, considering the high cost of clinical tests, long turnaround time, and lack of equal access around the world. Therefore, it is quite interesting and essential to distinguish COVID-19 patients from non-COVID-19 ones by evaluating their cough sound via a mobile application based on computer-aided digital technologies. In this way, the user can undergo constant self-surveillance wherever and whenever they want, which leads to infrequent medical visits as well as reducing the crowd in hospitals and the burden of healthcare teams. In this study, a method based on machine learning systems was presented to diagnose COVID-19 and non-COVID-19 patients with a single cough sound. The features were obtained by the MFCCs method from cough sounds and classified with seven different classifiers. The optimum hyperparameters of the system were selected according to the accuracy rate calculated with the LOO-CV strategy. In this way, COVID-19 and non-COVID-19 patients were classified with an accuracy rate of 0.9833, observing an increase in the accuracy of the most successful classifier (Euclidean-kNN) by around 7%. The system showed no error in the diagnosis of non-COVID-19 patients; however, it exhibited a sensitivity of 0.9720 for COVID-19 patients, which shows that it is quite successful compared with the systems available in the literature. Moreover, in this study, the used data set includes the unconscious and reflex cough sounds of COVID-19 patients. The results showed that conscious or unconscious cough sounds did not have any significance in the diagnosis of COVID-19. As mentioned, these systems can be installed on smartphones as an application or presented to users in a web-based environment. Today, mobile devices can be used as scalable, easy-to-use, and cost-effective health monitoring systems. Thus, for future studies, it is

planned to test the system on a larger number of samples and an online platform by adding serious preprocessing steps such as systems to distinguish cough sound from other sounds. In addition, only the PCR test was performed on the candidates in the data sets used in the study, and the label of the coughs was determined according to the result of this test. For this, unfortunately, comparing the results of the study with the results of different tests such as the antigen test is not possible. In the virufy dataset, there are people who are positive and do not have any symptoms. We believe that more comprehensive studies will emerge by considering these issues.

Appendix

See Figs. 3, 4, 5 and 6.

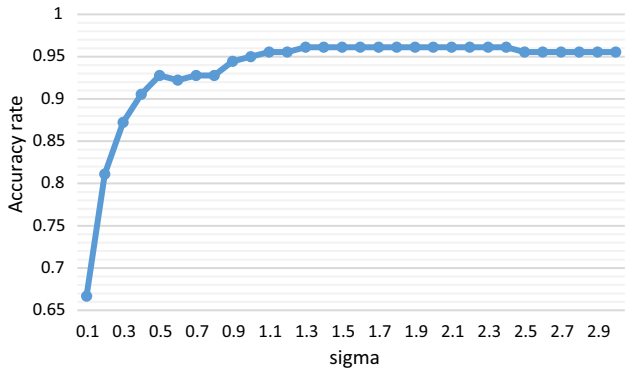

Fig. 3 The accuracy rate of the RBF-SVM classifier for different sigma values

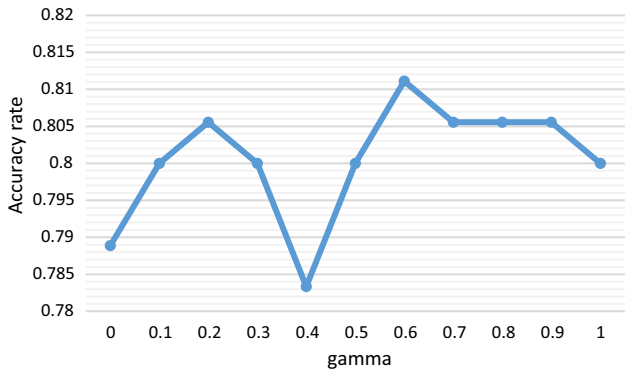

Fig. 4 The accuracy rate of the linear-LDA classifier for different gamma values

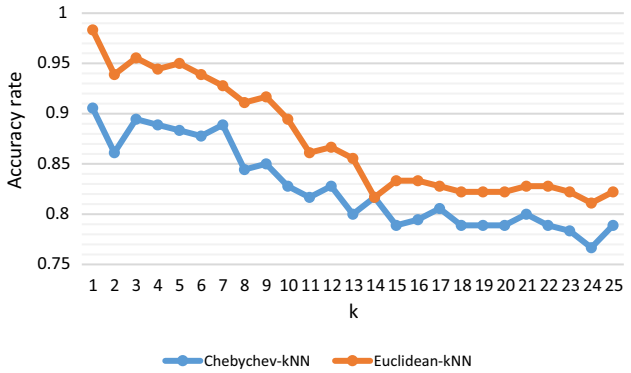

Fig. 5 The accuracy rate of Chebychev-kNN and Euclidean-kNN classifiers for different k values

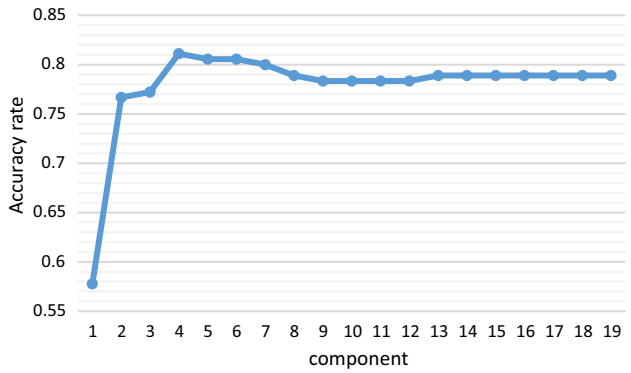

Fig. 6 The accuracy rate of the PLSR classifier for different components

Acknowledgment The author thanks all the healthcare teams for their hard work during the COVID-19 pandemic. Also, the author thanks Madison COHEN-MCFARLANE for sharing the NoCoCoDa data set.

Declarations

Conflict of interest The author does not have any financial and personal relationships with other people or organizations that could inappropriately influence his work.

References

1. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF (2020) The proximal origin of SARS-CoV-2. Nat Med 26(4):450–452. https://doi.org/10.1038/s41591-020-0820-9
2. "Coronavirus Disease (COVID-19) Situation Reports." https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports (accessed Feb. 06, 2021).
3. Wang D et al (2020) Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. JAMA—J Am Med Assoc 323(11):1061–1069. https://doi.org/10.1001/jama.2020.1585
4. Uysal C, Onat A, Filik T (2020) Non-contact respiratory rate estimation in real-time with modified joint unscented kalman

filter. IEEE Access 8:99445–99457. https://doi.org/10.1109/ACCESS.2020.2998117
5. Waheed A, Goyal M, Gupta D, Khanna A, Al-Turjman F, Pinheiro PR (2020) CovidGAN: data augmentation using auxiliary classifier GAN for improved covid-19 detection. IEEE Access 8:91916–91923. https://doi.org/10.1109/ACCESS.2020.2994762
6. Rajaraman S, Siegelman J, Alderson PO, Folio LS, Folio LR, Antani SK (2020) Iteratively pruned deep learning ensembles for COVID-19 detection in chest X-rays. IEEE Access 8:115041–115050. https://doi.org/10.1109/ACCESS.2020.3003810
7. Chamola V, Hassija V, Gupta V, Guizani M (2020) A comprehensive review of the COVID-19 pandemic and the role of IoT, drones, AI, blockchain, and 5G in managing its impact. IEEE Access 8:90225–90265. https://doi.org/10.1109/ACCESS.2020.2992341
8. Lan K, TongWang D, Fong S, Sheng Liu L, Wong KKL, Dey N (2018) A survey of data mining and deep learning in bioinformatics. J Med Syst 42(8):1–20. https://doi.org/10.1007/s10916-018-1003-9
9. Ali Humayun M et al (2019) Regularized urdu speech recognition with semi-supervised deep learning. Appl Sci 9(9):1956. https://doi.org/10.3390/app9091956
10. Shuja J, Alanazi E, Alasmary W, Alashaikh A (2020) COVID-19 open source data sets: a comprehensive survey. Appl Intell. https://doi.org/10.1007/s10489-020-01862-6
11. Srinivasa Rao ASR, Vazquez JA (2020) Identification of COVID-19 can be quicker through artificial intelligence framework using a mobile phone-based survey when cities and towns are under quarantine. Infect Control Hosp Epidemiol 41(7):826–830. https://doi.org/10.1017/ice.2020.61
12. Shi F et al (2020) Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for COVID-19. IEEE Rev Biomed Eng. https://doi.org/10.1109/RBME.2020.2987975
13. Sharma et al. N Coswara – A database of breathing, cough, and voice sounds for COVID-19 Diagnosis," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2020-October, pp. 4811–4815, May 2020, Accessed: Feb. 06, 2021. [Online]. Available: http://arxiv.org/abs/2005.10548.
14. Udugama B et al (2020) Diagnosing COVID-19: the disease and tools for detection. ACS nano 14(4):3822–3835. https://doi.org/10.1021/acsnano.0c02624
15. "Why your coronavirus test could cost $23—or $2,315." https://www.advisory.com/daily-briefing/2020/06/17/covid-test-cost (accessed Feb. 06, 2021).
16. Cohen-Mcfarlane M, Goubran R, Knoefel F (2020) Novel coronavirus cough database: NoCoCoDa. IEEE Access 8:154087–154094. https://doi.org/10.1109/ACCESS.2020.3018028
17. V. Bhateja, A. Taquee, and D. K. Sharma (2019) Pre-processing and classification of cough sounds in noisy environment using SVM. In: *2019 4th International Conference on Information Systems and Computer Networks, ISCON 2019*, pp. 822–826, doi: https://doi.org/10.1109/ISCON47742.2019.9036277
18. Simou N, Stefanakis N, Zervas P (2021) A universal system for cough detection in domestic acoustic environments. Eur Signal Process Conf 2021:111–115. https://doi.org/10.23919/Eusipco47968.2020.9287659
19. Barata F, Kipfer K, Weber M, Tinschert P, Fleisch E, Kowatsch T (2019) Towards device-agnostic mobile cough detection with convolutional neural networks. In: *2019 IEEE International Conference on Healthcare Informatics, ICHI 2019*, doi: https://doi.org/10.1109/ICHI.2019.8904554
20. Chatrzarrin H, Arcelus A, Goubran R, Knoefel F (2011) Feature extraction for the differentiation of dry and wet cough sounds. In: *MeMeA 2011 - 2011 IEEE International Symposium on Medical Measurements and Applications, Proceedings*, pp. 162–166, doi: https://doi.org/10.1109/MeMeA.2011.5966670
21. Nemati E, Rahman MM, Nathan V, Vatanparvar K, Kuang J (2020) A comprehensive approach for classification of the cough type. In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, vol. 2020-July, pp. 208–212, doi: https://doi.org/10.1109/EMBC44109.2020.9175345
22. Pramono RXA, Imtiaz SA, Rodriguez-Villegas E (2016) A cough-based algorithm for automatic diagnosis of pertussis. PLoS One 11(9):e0162128. https://doi.org/10.1371/journal.pone.0162128
23. Amoh J, Odame K (2016) Deep neural networks for identifying cough sounds. IEEE Trans Biomed Circuits Syst 10(5):1003–1011. https://doi.org/10.1109/TBCAS.2016.2598794
24. Cohen JP, Morrison P, Dao L, Roth K, Duong TQ, Ghassemi M (2021) COVID-19 image data collection prospective predictions are the future. J Mach Learn Biomed Imaging 2020:2–3
25. Cohen JP, Bertin P, Frappier V "Chester: A Web Delivered Locally Computed Chest X-Ray Disease Prediction System," *arXiv*, pp. 1–12, Jan. 2019, Accessed: Feb. 07, 2021. [Online]. Available: http://arxiv.org/abs/1901.11210.
26. Yang X, He X, Zhao J, Zhang Y, Zhang S, Xie P, "COVID-CT-dataset: A CT scan dataset about COVID-19," *arXiv*, Mar. 2020, Accessed: Feb. 07, 2021. [Online]. Available: http://arxiv.org/abs/2003.13865.
27. Wang S et al (2020) A deep learning algorithm using CT images to screen for corona virus disease COVID-19. MedRxiv. https://doi.org/10.1101/2020.02.14.20023028
28. Wang L, Lin ZQ, Wong A (2020) COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. Sci Rep 10(1):1–12. https://doi.org/10.1038/s41598-020-76550-z
29. Kleinberg B, van der Vegt I, Mozes M, Measuring Emotions in the COVID-19 Real World Worry Dataset, *arXiv*, Apr. 2020, Accessed: Feb. 07, 2021. [Online]. Available: http://arxiv.org/abs/2004.04225.
30. Banda JM et al, A large-scale COVID-19 Twitter chatter dataset for open scientific research – an international collaboration, *arXiv*, Apr. 2020, Accessed: Feb. 07, 2021. [Online]. Available: http://arxiv.org/abs/2004.03688.
31. Chen E, Lerman K, Ferrara E (2020) Tracking social media discourse about the COVID-19 pandemic: development of a public coronavirus Twitter data set. JMIR Public Heal Surveill 6(2):e19273. https://doi.org/10.2196/19273
32. Huang Y et al (2020) The respiratory sound features of COVID-19 patients fill gaps between clinical data and screening methods. MedRxiv. https://doi.org/10.1101/2020.04.07.20051060
33. Quatieri TF, Talkar T, Palmer JS (2020) A framework for biomarkers of COVID-19 based on coordination of speech-production subsystems. IEEE Open J Eng Med Biol 1:203–206. https://doi.org/10.1109/ojemb.2020.2998051
34. Han J et al. An Early Study on Intelligent Analysis of Speech under COVID-19: Severity, Sleep Quality, Fatigue, and Anxiety, *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 2020-October, pp. 4946–4950, Apr. 2020, Accessed: May 30, 2021. [Online]. Available: http://arxiv.org/abs/2005.00096.
35. Alsabek MB, Shahin I Hassan A (2020) Studying the similarity of COVID-19 sounds based on correlation analysis of MFCC. In: *Proceedings of the 2020 IEEE International Conference on Communications, Computing, Cybersecurity, and Informatics, CCCI 2020*, Nov. 2020, doi: https://doi.org/10.1109/CCCI49893.2020.9256700.
36. Brown C et al (2020) Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data. Proc ACM

SIGKDD Int Conf Knowl Discov Data Min 11:3474–3484. https://doi.org/10.1145/3394486.3412865
37. Imran A et al (2020) AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. Inform Med Unlocked 20:100378. https://doi.org/10.1016/j.imu.2020.100378
38. Feng K, He F, Steinmann J, Demirkiran I (2021) Deep-learning based approach to identify covid-19. In: *Conference Proceedings - IEEE SOUTHEASTCON*, vol. 2021-March, doi: https://doi.org/10.1109/SoutheastCon45413.2021.9401826.
39. Vijayakumar DS, Sneha M (2021) Low cost Covid-19 preliminary diagnosis utilizing cough samples and keenly intellective deep learning approaches. Alexandria Eng J 60(1):549–557. https://doi.org/10.1016/j.aej.2020.09.032
40. Pal A, Sankarasubbu M (2021) Pay attention to the cough: early diagnosis of COVID-19 using interpretable symptoms embeddings with cough sound signal processing. In: *Proceedings of the ACM Symposium on Applied Computing*, pp. 620–628, doi: https://doi.org/10.1145/3412841.3441943.
41. Bagad et al. P Cough against COVID: evidence of COVID-19 signature in cough sounds. *arXiv*, Sep. 2020, Accessed: Feb. 06, 2021. [Online]. Available: http://arxiv.org/abs/2009.08790.
42. Manshouri N (2021) Identifying COVID-19 by using spectral analysis of cough recordings: a distinctive classification study identifying COVID-19 by using spectral analysis of cough recordings: A distinctive classification study Negin MANSHOURI," Preprints, Accessed: Feb. 06, 2021. [Online]. Available: www.preprints.org.
43. Chaudhari G et al. (2020) Virufy: global applicability of crowdsourced and clinical datasets for AI detection of COVID-19 from cough, *arXiv*, Nov., Accessed: Feb. 06, 2021. [Online]. Available: http://arxiv.org/abs/2011.13320.
44. Han W, Chan CF, Choy CS, Pun KP (2006) An efficient MFCC extraction method in speech recognition. In: *Proceedings—IEEE International Symposium on Circuits and Systems*, pp. 145–148, doi: https://doi.org/10.1109/iscas.2006.1692543.
45. Schafer RW, Rabiner LR (1975) Digital representations of speech signals. Proc IEEE 63(4):662–677. https://doi.org/10.1109/PROC.1975.9799
46. Atal BS (1976) Automatic recognition of speakers from their voices. Proc. IEEE 64(4):460–475. https://doi.org/10.1109/PROC.1976.10155
47. Patel K, Prasad RK (2013) Speech recognition and verification using MFCC & VQ. Int J Emerg Sci Eng (IJESE) 1(7):33–37
48. Davis SB, Mermelstein P (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust Speech Signal Process 28(4):357–366. https://doi.org/10.1109/TASSP.1980.1163420
49. Bost R, Popa RA, Tu S, Goldwasser S (2015) Machine learning classification over encrypted data, network and distributed system security symposium. https://doi.org/10.14722/ndss.2015.23241
50. Melek M, Manshouri N, Kayikcioglu T (2020) Low-cost brain-computer interface using the emotiv epoc headset based on rotating vanes. Trait du Signal 37(5):831–837. https://doi.org/10.18280/ts.370516
51. Pahar M, Klopper M, Warren R, Niesler T (2020) COVID-19 cough classification using machine learning and global smartphone recordings, Accessed: Feb. 06, 2021. [Online]. Available: http://arxiv.org/abs/2012.01926.
52. Melek M, Manshouri N, Kayikcioglu T (2020) An automatic EEG-based sleep staging system with introducing NAoSP and NAoGP as new metrics for sleep staging systems. Cogn Neurodyn. https://doi.org/10.1007/s11571-020-09641-2
53. Webb GI et al. Leave-one-out cross-validation. In: *Encyclopedia of Machine Learning*, Springer US, 2011, pp. 600–601
54. "iOS App Store Requirements For Health Apps | Dash Solutions Blog." https://blog.dashsdk.com/app-store-requirements-for-health-apps/ (accessed Jul. 05, 2021).
55. "Google Play Store Requirements For Health Apps | Dash Solutions Blog." https://blog.dashsdk.com/play-store-requirements-for-health-apps/ (accessed Jul. 05, 2021).
56. Raudys SJ, Jain AK (1990) Small sample size effects in statistical pattern recognition: recommendations for practitioners and open problems. Proc—Int Conf Pattern Recognit 1:417–423. https://doi.org/10.1109/icpr.1990.118138