A Novel Approach to COVID-19 Diagnosis Based on Mel Spectrogram Features and Artificial Intelligence Techniques

Aseel Alfaidi†, Abdullah Alshahrani†, and Maha Aljohani††
aalfaidi0005.stu@uj.edu.sa asalshahrani2@uj.edu.sa mmaljohani@uj.edu.sa
†Department of Computer Science and Artificial Intelligence, College of Computer Science and Engineering, University of Jeddah, Jeddah 21493, Saudi Arabia
††Department of Software Engineering, College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia

Abstract
COVID-19 has remained one of the most serious health crises in recent history, resulting in the tragic loss of lives and significant economic impacts on the entire world. The difficulty of controlling COVID-19 poses a threat to the global health sector. Considering that Artificial Intelligence (AI) has contributed to improving research methods and solving problems facing diverse fields of study, AI algorithms have also proven effective in disease detection and early diagnosis. Specifically, acoustic features offer a promising prospect for the early detection of respiratory diseases. Motivated by these observations, this study conceptualized a speech-based diagnostic model to aid in COVID-19 diagnosis. The proposed methodology uses speech signals from confirmed positive and negative cases of COVID-19 to extract features through the pre-trained Visual Geometry Group (VGG-16) model based on Mel spectrogram images. This is used in addition to the K-means algorithm that determines effective features, followed by a Genetic Algorithm-Support Vector Machine (GA-SVM) classifier to classify cases. The experimental findings indicate the proposed methodology's capability to classify COVID-19 and NOT COVID-19 of varying ages and speaking different languages, as demonstrated in the simulations. The proposed methodology depends on deep features, followed by the dimension reduction technique for features to detect COVID-19. As a result, it produces better and more consistent performance than handcrafted features used in previous studies.

Keywords:
Artificial Intelligence, COVID-19 diagnosis, Speech signals, Mel spectrogram features, Transfer learning

1. Introduction

The difficulty of controlling the coronavirus (COVID-19) is considered a threat to global public health. COVID-19 was discovered in late 2019 in Wuhan, China, and has since spread worldwide. In March 2020, the World Health Organization (WHO) declared that the COVID-19 outbreak had become a pandemic [1]. The pandemic has negative impacts on society, health, and the economy. Two other issues of concern are the continued spread of the virus and its multiple strains that all exhibit the same symptoms.

To date, the number of COVID-19 cases globally has exceeded more than 500 million, increasing the demand for screening, diagnosis, and testing of individuals. One of the primary methods of detecting COVID-19 is testing for Reverse Transcription-Polymerase Chain Reaction (RT-PCR) by detecting the presence of viral ribonucleic acid (RNA) from swab samples [2]. However, this method is insufficient to fight the pandemic for several reasons. First, the test takes up to several days, and the length of time varies by country. Second, the test requires visiting clinics, and if proper precautions are not taken, this will expose many people or medical staff to COVID-19. The third issue involves the high cost and scarcity of this test in some countries.

One of the examination methods used, the Rapid Antigen Test (RAT), is a common alternative to RT-PCR testing [3]. It is less expensive and takes a shorter time than RT-PCR testing. However, samples still have to be taken in clinics, posing the same limitation as that of the first method.

Progress in Artificial Intelligence (AI) has contributed to improving and solving problems facing all fields worldwide. The fields of AI and Big Data (BD) can play a significant role in healthcare by detecting and tracking the growth rate of COVID-19 [4] [5]. Additionally, it supports patient care through early diagnosis and monitoring methods. Thus, an AI-based examination's applicability offers a high potential for COVID-19 in terms of patient status tracking, prompt result processing, and reduced spread, all at a low cost.

Previous studies analyzed the pathological changes caused by COVID-19 in the respiratory system and revealed some patterns in the vocal cords and the intensity of the voice change in people infected with the COVID-19 virus [6] [7]. In accordance with these findings, in the current study, we aim to extend the idea with an alternative speech-based diagnosis approach. As a result, the main contributions of this research are as follows:

- Using speech signals from COVID-19 patients as an alternative method of diagnosis;
- Applying the transfer learning approach through the pre-trained ImageNet model on audio datasets

Manuscript received September 5, 2022
Manuscript revised September 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.9.29

to extract the most accurate and powerful features;
- Proposing a pre-trained Visual Geometry Group (VGG-16) model for extracting the deep features based on Mel spectrogram images.
- Using K-means algorithms to determine effective features from the deep features;
- Combining the Genetic Algorithm (GA) and the Support Vector Machine (SVM) algorithm to classify cases;
- Assessing the applicability of speech-based diagnosis to COVID-19 for different age groups, genders, and languages; and
- Measuring the efficiency of the speech-based diagnosis of COVID-19 by considering the patient's symptoms, pre-existing diseases, and smoking habit.

We have organized the remainder of this paper as follows. In Section 2, we review the literature on disease diagnosis and COVID-19 through speech signals. In Section 3, we describe the dataset, as well as the data processing steps. In Section 4, we explain our proposed methodology and algorithms for diagnosing COVID-19. In Section 5, we present the experimental results, followed by a discussion. Finally, in Section 6, we draw our conclusion.

2. Literature Review

Speech is a non-invasive biomarker that has been used to assess human body pain [8] and detect depression [9]. It is also a diagnostic tool for various diseases, such as vocal disorders [10], multiple sclerosis [11], Parkinson's disease [12], and heart failure [13]. In this context, Schuller et al. [14] provided an overview of computer audition (CA), such as audio analysis by AI, to aid in controlling the COVID-19 pandemic in various ways, including risk assessment, diagnosis, and monitoring of its spread. Recent studies have focused on COVID-19 diagnosis based on cough and breathing sounds [15] [16] [17]. A relatively small group of publications expressed interest in speech.

For instance, Han et al. [18] conducted a preliminary study of COVID-19 patient speech analysis to categorize the patients' health status based on disease severity, sleep quality, fatigue, and anxiety. The experiments involved 51 people infected with COVID-19 in two hospitals in Wuhan. The authors used the Computational Paralinguistics Challenge (ComParE) set and the Extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) as audio features, as well as the SVM classifier. The SVM classifier obtained an accuracy of 69% in estimating the severity of the disease.

Verde et al. [19] presented a study using Machine Learning (ML) algorithms and the Coswara dataset to detect COVID-19 by analyzing speech signals. The evaluated dataset consisted of 83 healthy and 83 COVID-19 pathological cases. As audio features, Fundamental Frequency (F0), shimmer, jitter, and Harmonics-to-Noise-Ratio (HNR) were used. This study showed that the Random Forest algorithm could classify and achieve up to 85% accuracy when using the vowel e for diagnosis.

Usman et al. [20] discussed COVID-19 detection from speech data using spectral features and ML algorithms. Their study was based on the speech of 22 patients infected with COVID-19 and 84 healthy subjects. The authors noted that all classification algorithms achieved around 70% in recall value, and the best performance of Decision Forest (DF) algorithms achieved an accuracy of 78%.

In contrast, symptoms and speech were also explored in a study by Han et al. [21] to discriminate between COVID-19-positive cases and healthy cases. According to the COVID-19 sounds dataset, the study was conducted on 362 positive cases, 502 negative cases, and symptoms. This study used an acoustic feature (ComParE) set and an SVM classifier for diagnosis and achieved an area under the curve (AUC) value of 0.79.

Additionally, a study was conducted by Stasak et al. [22] to detect COVID-19. They used the speech of participants with COVID-19 symptoms and similar symptoms with a positive or a negative result. The dataset included 44 healthy participants, 22 COVID-19-positive participants, and symptoms. The authors used the glottal, prosodic, and spectral features from the A Cooperative Voice Analysis Repository for Speech Technologies (COVAREP) features and the Decision Tree (DT) classifier. The results indicated that speech features with symptoms may produce a COVID-19 classification accuracy of up to 80%.

The existing literature [19] [20] [21] [22] emphasizes that COVID-19 affects the vocal tract and that acoustic analysis can detect and diagnose COVID-19. Thus, this confirms one of our objectives of analyzing speech signals to diagnose COVID-19. Table 1 presents the techniques and datasets used in the literature related to COVID-19 and the performance results. The accuracy rates obtained are still relatively low and can be further improved. Extracting features through handcrafted methods is a common strategy used in these studies.

In comparison, many Deep Learning (DL) technologies have been developed to extract features and improve performance, where selecting features has a significant impact and may lead to correct discrimination between classes and optimal performance of the classifier. DL uses neural networks inspired by a human brain's structure, which can learn and analyze the relations among the data to extract features. Recently, DL techniques have been

used to extract deep features across various datasets in the audio domain for diagnoses of diseases and have achieved superior performance compared with traditional methods.

For example, García-Ordás et al. [23] proposed the detection of respiratory pathologies through breaths using the Convolutional Neural Network(CNN) model and the Mel spectrogram. Their study indicated that the model achieved up to 99% sensitivity and 99% specificity. Also, Vavrek et al. [24] detected dysphonia pathologies using the pre-trained ImageNet VGG-16 and a spectrogram for speech signals. The study explained can be used for the DL technique for voice signals and achieved an accuracy of up to 82%.

Zahid et al. [12] suggested Parkinson's disease diagnosis using the pre-trained ImageNet (Alexnet) and a spectrogram through speech. The authors reported that the deep feature performed better than systems based on simple acoustic properties and achieved an accuracy of up to 99.1%. Finally, Zhou et al. [25] proposed cough recognition for COVID-19 through the CNN model and the Mel spectrogram. The proposed approach to diagnosis achieved an accuracy of up 98%.

DL models have demonstrated their ability to learn features from spectrograms for many tasks and have the ability to adequately capture the variability among features, unlike handcrafted features. Thus, extracting the most accurate and powerful features is the optimization goal of the target task. As a result, in this study, we apply the pre-trained (VGG-16) model, which might extract more valuable features from the speech signals, and the SVM algorithm to classify the COVID-19 cases.

Table 1 : Brief Overview of Studies Using Speech Data to Diagnose COVID-19

Authors	Data	Feature	Model	Performance Metrics
[18]	Sample audio from two hospitals in Wuhan	ComParE and eGeMAPS acoustic feature set	SVM	Accuracy 67%
[19]	Coswara	F0, Jitter, Shimmer, HNR	Random forest	Accuracy 85%
[20]	Not mentioned	Spectral features	DF	Accuracy 73%
[21]	COVID-19 sounds	ComParE acoustic feature set	SVM	AUC 0.79
[22]	Sonde Health COVID-19	COVAREP acoustic feature set	DT	Accuracy 80%

3. Data

In this section, we provide a detailed overview of the dataset and the steps of processing the dataset.

3.1 Data Description

The Department of Computer Science and Technology at the University of Cambridge [26] has created a crowdsourced database of sounds that has been compiled from various sources to be used for COVID-19. It contains sound samples of coughing, breathing, and reading a short audio voice ("I hope my data will be useful in managing the virus pandemic.") to aid in the diagnosis of the virus infection. It includes audio samples from both the positive and the negative cases, presented by the participants. It also contains demographic information about each participant, such as age, gender, language, symptoms, and medical history. While speech signals convey information about human health, they have also been used as a tool for assessing and diagnosing a wide range of diseases. Our study therefore relies solely on speech cues to diagnose COVID-19.

The crowdsourced dataset contains many participants' cases to ensure the effectiveness of the diagnostic process. We therefore chose participants who confirmed their COVID-19 results as positive or negative. Additionally, whereas the data was gathered through crowdsourcing, it was manually checked for audio quality, and any corrupted audio data with poor quality was discarded. As a result, 623 COVID-19-positive participants and 774 COVID-19-negative participants were selected. The demographic characteristics of the samples used in this study are depicted in Figure 1.

3.2 Data Processing

In the audio processing step, we standardized the sampling rate of the audio signals at 16 kHz, so all arrays had the same dimensions. We also removed the silent periods in the beginning and at the end of the audio. Then, we resized the audio to the same length by dividing it into 5-second segments.

3.3 Data Augmentation

Data augmentation is a method of dealing with the problem of having insufficient training data. It comprises data that has been generated from an existing data set using a variety of techniques. Adding background noise to

an audio signal is one method of enhancing the data contained within the audio signal.

To generate noise, we used the Additive White Gaussian Noise (AWGN) approach, which uses a Gaussian distribution with a mean of zero and a standard deviation equal to the root mean square (RMS) of the value of the noise, which was calculated using Equation 1:

$$RMS_{notse} = \sqrt{\frac{RMS^2{}_{signal}}{10^{\frac{SNR}{10}}}} \quad (1)$$

where RMS refers to the root mean square of the value of a signal, and SNR represents the signal-to-noise ratio = 10.

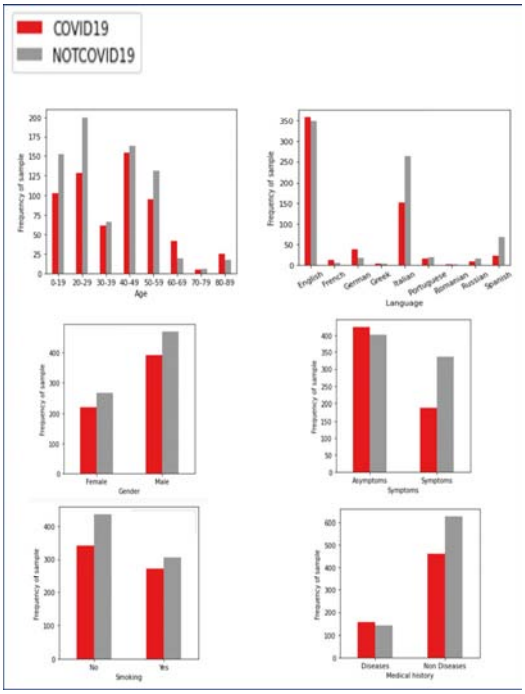

Fig. 1 Overview of Sample Distribution

4. Research Methodology

To bridge the gap in the existing methods of COVID-19 diagnosis, we concentrate on analyzing speech signals as an alternative method of diagnosis.

The proposed methodology comprises four main phases: the audio processor, the pre-trained (VGG-16) model, dimension reduction by the k-means algorithm, and the GA-SVM classifier, as illustrated in Figure 2.

4.1 Audio Processor

During the first phase, after all of the audio data had been processed, the data used in this study totaled 3829 cases (2376 positive and 1453 negative). We converted all the audio data into the Mel spectrogram images, which allowed extracting features from the pre-trained (VGG-16) model.

The Mel spectrogram is a technology for studying audio, in which frequencies are converted into the Mel scale. The Mel scale is intended to mimic the way that the human auditory system operates, where discrimination is greater for lower frequency sounds than for higher frequency sounds [27]. It is also a tool for visualizing the change in frequency of an audio signal over time [28].

Additional benefits of the Mel spectrogram are its visually appealing way of representing the signal strength of an audio and its effectiveness as a tool for extracting hidden features from an audio. The Mel spectrogram also stores detailed information that allows the appearance of differentiation in the audio shape. In the case of audio feature extraction, the Mel spectrogram techniques have been extensively used in previous research [9] [23] [25] [29]. The computation of a Fast Fourier Transform (FFT) and a Mel-filter bank are the foundations of the Mel spectrogram [27]. The FFT, which represents a signal conversion from the time domain to the frequency domain, is calculated using Equation 2:

$$X(k) = \sum_{n=0}^{N-1} X(n) \times e^{-j\frac{2\pi}{N}} \times kn \quad k = 0,1,2\,I.,N-1 \quad (2)$$

where X(k) denotes the frequency domain sample, X(n) represents the time domain sample, and N represents the FFT size.

A Mel-filter bank converts the frequency to the Mel scale, is calculated using Equation 3:

$$Mel = 1127 \times log(\frac{frequency}{700} + 1) \quad (3)$$

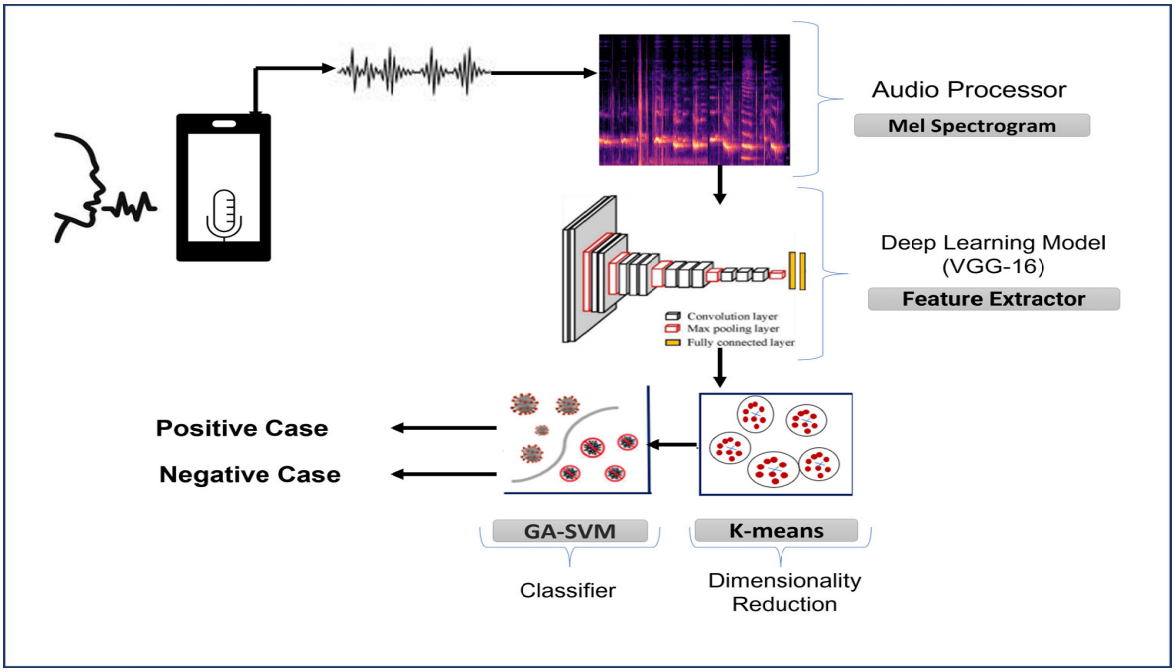

Fig.2 Proposed Methodology for Diagnosing COVID-19

The Mel spectrogram was created by splitting the signal into short static frames, with a length of 25 milliseconds and an interval of 10 milliseconds between frames. Then, for each frame, the FFT with a length of 1024 was applied to obtain the time frequency for that particular frame. Once this process was completed, each frequency-domain frame was passed through the Mel-filter bank to convert it to the Mel scale. In the last step, the Mel spectrogram was created by adding the results of each Mel filter. Librosa [30], a Python package for audio analysis and processing, was used in this study to create a Mel spectrogram. The steps involved in extracting the Mel spectrogram from an audio signal are depicted in Figure 3.

Figure 4 also depicts the Mel spectrogram of a COVID-19 carrier and a voice that has not been infected with the virus (not infected). It also demonstrates the difference in tones between the two audios. Variations in hue indicate the signal strength in the Mel spectrogram, with the light color indicating speech events and high energy and the dark color indicating a break in speech and low energy. The positive case shows a break in the voice and low energy when speaking, while the negative case shows no break in the voice and high energy when speaking.

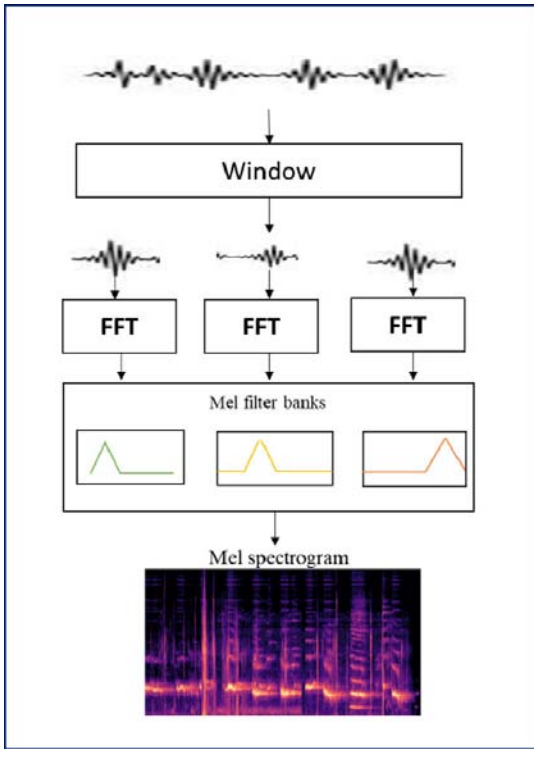

Fig.3 Steps of Converting Audio Signals into the Mel Spectrogram

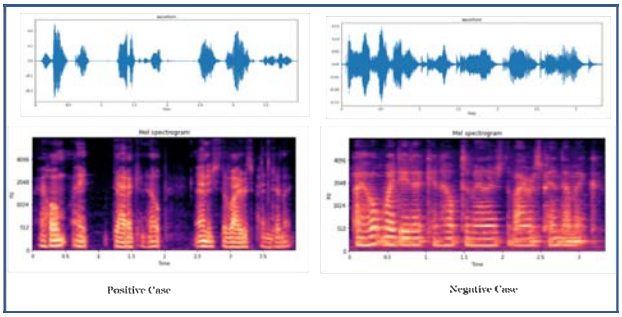

Fig.4 Mel Spectrogram Samples of Positive-/Negative-Case Speech Signals

4.2 Deep Learning Model

The transfer learning approach was used in the course of the work in the second phase. We used the pre-trained (VGG-16) network proposed by Simonyan and Zisserman [31] as a feature extractor from the Mel spectrogram. The VGG-16 architecture is generated by stacking 3 x 3-filter size convolutional (CONV) layers with maximum (Max)-pooling layers for the 2 x 2-filter size. Because the dimensions of the VGG-16 model decrease, while the depth increases with each layer, the model's strength is attributed to this property. The pre-trained VGG-16 model is described in detail in Figure 5, and the layers used in the research are explained in greater detail.

1. Input Layer

The input layer of the VGG-16 model uses images that have three channels and 224 x 224 x 3 sizes. Therefore, we resized the Mel spectrogram images to be proportional to the model as inputs.

2. Feature extraction layer

The feature extraction layer is made up of the CONV layer, the Max-pooling layer, and two fully connected (FC) layers, as illustrated in Figure 5. Therefore, we passed the Mel spectrogram images through these layers to extract the deep features. The following are the layers' specific characteristics:

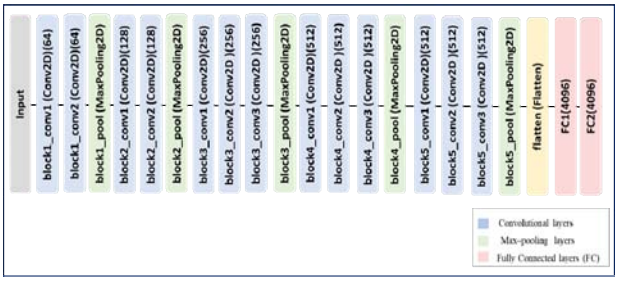

Fig.5 VGG-16 Model of Proposed Methodology

- **CONV layer**. It is the fundamental component of the model and the first layer. This layer functions as a window to search for features in the input images, pixel by pixel, to extract the features from the images. Additionally, the convolution process is carried out in this layer between the input image and the convolutional filter, resulting in the production of a feature map, which is a representation of the image, with new pixel values derived from the original image [32].

- **Max-pooling layer.** This layer contributes to the reduction in the size of the network by decreasing the number of parameters provided to the next layer [32]. As a result, the max-pooling layer contains only the essential features from the previous feature map. The VGG-16 model has five max-pooling layers, following the CONV layers, as shown in Figure 5.

- **FC layers.** These final layers in the model gather the features from the previous layers and perform a high-level logical action between them. Although there are three FC layers in a VGG-16 model, the last layer (FC3) is the one that is used for classification. As a result, in this study, we used two FC layers with a total of 4,096 nodes to extract deep features, as illustrated in Figure 5.

4.3 Dimensionality Reduction

In the third phase, we aim to have the most appropriate and relevant features that would help in COVID-19 diagnosis. Therefore, the model relies on the k-means algorithm, which reduces the deep features extracted from VGG-16 before applying the GA-SVM classifier.

Following the feature extraction step from the VGG-16 model, we obtained 4096 dimensions for each instance in the dataset. To map relevant features from the original features had been reduced to a smaller number of features [33]. In data mining, the phrase "reduce dimensions" means reducing the number of features in a dataset, while

retaining the greatest amount of variance from the original dataset. As a result, it determines which features are the most important and improves the performance of ML algorithms [34].

The clustering algorithm is used to reduce dimensions. It is an unsupervised learning technique, defined as the process of grouping data points into a number of groups in such a way that the data points from the same groups are more closely related to one another than the data points from other groups [33]. As a dimensionality reduction technique, it is defined as the process of grouping data points into a number of clusters and computing the distance of a data point from each cluster center by the Euclidean Distance function using Equation 4. Finally, it represents each of the data points in terms of how far it is from each of these cluster centers as the feature vectors for data [33] [35].

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \| x_i^{(j)} - c_j \|^2 \quad (4)$$

where k represents the number of clusters, n denotes the dataset points, and c signifies the centroid for the cluster. The distance indicator of the points $x_i^{(j)}$ and their represented centroids c_j are $\| x_i^{(j)} - c_j \|$.

The k-means algorithm is a type of clustering algorithm that is most commonly used to reduce dimensions [36] [37] [38]. This algorithm fundamentally depends on the number of clusters; consequently, the elbow method [33] is used to determine the optimal number of clusters to use. The elbow method is represented by a graph that depicts the sum of squared errors (SSE) for each value of the parameter k. The SSE is defined as the sum of the squared distances between each data point in the cluster and the centroid of the cluster, and it is calculated using Equation 5.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} distance(m_i, x)^2 \quad (5)$$

where x represents the data point in each cluster C_i, and m_i is the representative for the center cluster C_i.

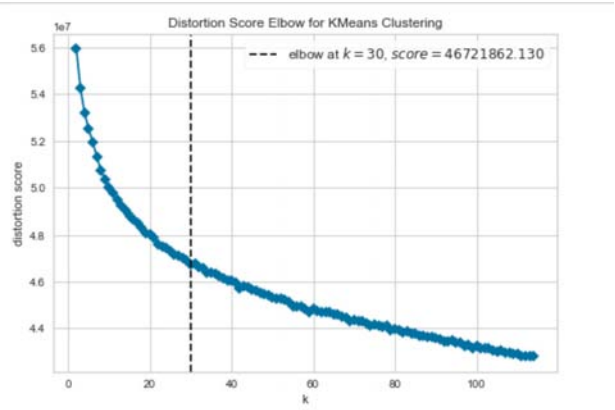

Fig.6 Elbow Method for K Value

Figure 6 illustrates the elbow method graph when applied to our dataset. First, we randomly initialized the numbers of the k cluster and displayed them against the SSE. As shown in Figure 6, k = 30 is the elbow point, and it is the optimal k value in our dataset. After determining the k value for the clusters, we applied the k-means algorithm to the dataset by calculating the distance of each data point from each centroid using Equation 4.

4.4 Classifier

After obtaining the audio features from the previous stages, it is fed into the GA-SVM classifier, which classifies the case as either COVID-19 or NOT COVID-19.

The SVM algorithm is a supervised learning algorithm for classification [39]. It is also a powerful and adaptable tool, capable of implementing linear or nonlinear classification, regression, and even outlier detection [33]. Figure 7 illustrates the basic concept of the SVM, which is based on finding the best way to divide data through a hyperplane. Support vectors are used to select the optimal hyperplane, which is a collection of data points that are closest to the hyperplane.

Basically, the hyperplane is calculated using Equation 6, where w is an n-dimensional vector, x represents the input feature vector, and b denotes the bias.

$$w^T.x + b = 0 \quad (6)$$

This divides the data into two classes as the label can be either '+1' for Class A or '-1' for Class B, as shown in Figure 7. The data is labeled in the classifier based on the conditions, as indicated in Equations 7 and 8.

$$if\ w^T.x + b \geq 0, y = 1 \quad (7)$$

$$if\ w^T.x + b \leq 0, y = -1 \quad (8)$$

Additionally, Figure 7 illustrates the margin and is calculated using Equation 9. It is as the distance between the hyperplane and the support vector for each class. The greater the distance, the higher the probability that new data would be correctly classified [40].

$$d(w,b;x) = \frac{|(w^T.x + b - 1) - (w^T.x + b + 1)|}{||w||} = \frac{2}{||w||} \quad (9)$$

The SVM is one of the most commonly used and popular classification algorithms for disease diagnosis [36] [38] [41]. Furthermore, it is highly effective in classification and has a high degree of generalization capacity. In general, the performance of the SVM classifier depends on the chosen hyperparameters, such as kernel functions and the parameter C [40]. C is a regularization and optimization parameter for maximizing the margin and minimizing the classification error. The kernel is another parameter that receives data as input and converts it into the format required for data processing.

For the SVM's best performance, hyperparameter adjustment is essential. However, there is no precise benchmark for evaluating the value of each SVM hyperparameter that may be used. As a result, optimized algorithms, such as GA, provide tuning for hyperparameters. Genetic-based optimization is a technique that uses genetics and natural selection to find the best solution. It is based on Darwin's theory of evolution and implies the (survival of the fittest) [42].

Figure 8 presents how in this study, we used GA for hyperparameter optimization of the SVM. The GA process begins with a population of randomly generated candidate solutions. It is a set of chromosomes evaluated by the fitness function and includes the accuracy rate. At the same time, each chromosome has a hyperparameter and the actual input value for each evaluation. Iterative searching for high-performing hyperparameters combines each generation and passes them on to the next until the highest-performing combination has been discovered. Although parameter values are initially generated at random, they rarely contain optimal parameter values. The genetic factors of selection, crossover, and mutation are used by the algorithm to determine these parameters [42]. The following list provides an overview of genetic operators:

- Selection operator. The concept of selection operation is to choose the hyperparameter with the highest fitness value to be used in the next generation of the algorithm.
- Crossover operator. This is a process of chromosome interchange that results in the creation of a new individual, and we used the uniform crossover to accomplish this. It indicates that each gene is determined independently by a 50% random distribution of chromosomes.
- Mutation operator. The mutation operator is used in the process of generating a new generation. Mutation represents the introduction of new patterns into the chromosomes, as well as the random modification of the information contained within.

The loop is repeated until a stopping condition is met. A condition has been established for stopping when the maximum number of generations has been reached. At that point, the optimal SVM hyperparameter has been determined, and the model has been tested on the data.

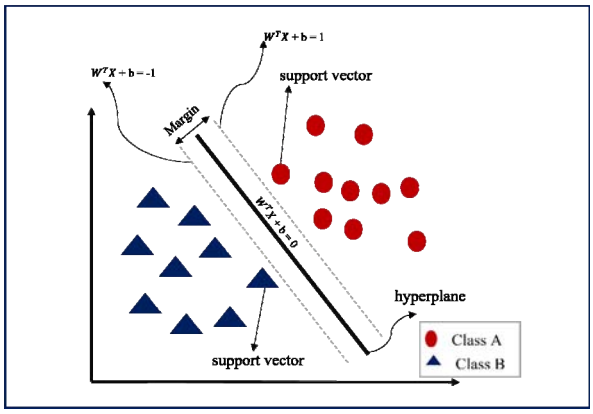

Fig.7 Support Vector Machine Algorithm

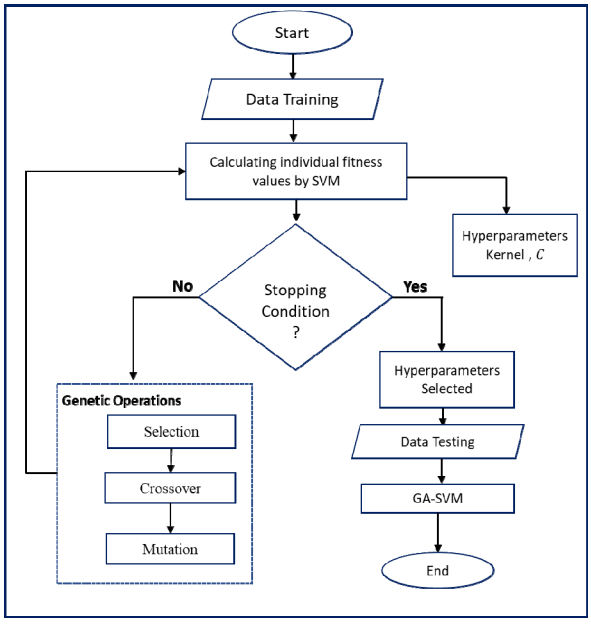

Fig. 8 Flowchart Diagram of GA with SVM

5 Results and Discussion

In this section, we present our study's results and evaluate the model in a variety of scenarios, as well as discuss the specifics of these results. Following this, we demonstrate and compare the performance of our model with those of other models that use the same or different phenomena.

5.1 Performance Metrics

Specifically, the ramifications of misclassifying a sample as not being infected with COVID-19 may cause the spread of the infection. Consequently, the model's performance is measured using more performance metrics, as described below.

Accuracy is the percentage of correct predictions made when testing a model and is calculated in Equation 10.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FN+FP)} \quad (10)$$

Sensitivity is the percentage of times that a model correctly predicts positive results and is calculated in Equation 11.

$$Sensitivity = \frac{TP}{(TP+FN)} \quad (11)$$

Finally, specificity is the percentage of times that a model correctly predicts the negative results and is calculated in Equation 12.

$$Specificity = \frac{TN}{(TN+FP)} \quad (12)$$

To explain in more detail the variables used in performance metrics equations, true positive (TP) is the number of samples infected with COVID-19 that the model actually predicted as infected. In contrast, true negative (TN) represents the number of NOT COVID-19 samples that the model predicted as not infected. In contrast, false positive (FP) and false negative (FN) variables represent the number of COVID-19 or NOT COVID-19 samples, respectively, and were incorrectly predicted by the model.

The receiver operating characteristic (ROC) curve was also used to assess the model's performance. It is a graph that illustrates the relation between the TP and the FP rates. More specifically, it displays the AUC of the data. It will be more accurate for the classifier to distinguish between positive and negative class points if the value is closer to one.

5.2 Experimental Results

To know the diagnostic efficiency of the speech signals for COVID-19, we performed two experiments for the evaluation. Both experiments were conducted using data divided 80/20, with 80% of the data used for training and 20% for testing.

5.2.1 First Experiment

The model was evaluated when conducting the first experiment by dividing the data into similar groups based on age, language, and gender. Additional evaluations included medical history, smoking status, and COVID-19 symptoms of the samples.

To take into account multilingualism and the spread of COVID-19 worldwide, we investigated the relation between the human voice and COVID-19 infection, and COVID-19 was diagnosed in each language separately. It turned out that the model's accuracy rates were 99% in English, 98% in Italian, and 97% in Spanish. In comparison, its accuracy rates were 93% and 92% in Russian and French, respectively, 95% in Portuguese and German, and 100% in Greek and Romanian.

All people of all ages are at risk of contracting COVID-19, so we evaluated the model based on age groups at 10-year intervals. The results showed the following diagnostic accuracy rates: 93% for the 20–29 age group, 95% each for the 0–19 and 50–59 age groups, 96% each for the 40–49, 60–69, and 80–89 age groups, and 97% and 100% for the 30–39 and 79–70 age groups, respectively. Additionally, the model reported an equal diagnostic accuracy of 96% for both genders.

Those who experience symptoms, such as coughing and sore throat, may have more negative impacts on their respiratory systems than those who do not experience symptoms. Therefore, we isolated those experiencing symptoms from either the positive or the negative test. At the same time, we separated those who did not show any signs of illness. When we tested the model, it achieved a similar accuracy of 97% for both groups with and without symptoms.

To study the accuracy of the COVID-19 diagnosis based on speech signals from those who had respiratory or other diseases, we examined the responses of those who reported having diseases and those who had no medical history of diseases. The model has the ability to perform COVID-19 diagnosis for those who have diseases, with an accuracy of 96%. Furthermore, the model achieved a 99% accuracy rate for those who did not have any diseases.

Smoking is also a factor that has adverse impacts on the respiratory system and the vocal cords of an individual. Therefore, to determine the model's effectiveness in diagnosing this category based on the audio characteristics, we evaluated the model through samples of smokers and

Table 2 Performance Metrics of GA-SVM Classifier

Group	Accuracy (%)	Sensitivity (%)	Specificity (%)
First Experiment			
a. Language			
English	99	99	100
Italian	98	95	99
Spanish	97	95	98
Portuguese	95	90	89
Russian	93	88	100
French	92	92	100
German	95	95	92
Greek	100	100	100
Romanian	100	100	100
b. Age range			
0–19	95	95	95
20–29	93	95	89
30–39	97	98	93
40–49	96	94	90
50–59	95	99	97
60–69	96	95	100
70–79	100	100	100
80–89	96	95	100
c. Gender			
Female	96	96	97
Male	96	96	97
d. Symptoms of COVID-19			
Symptomatic	97	92	99
Asymptomatic	97	98	94
e. Medical history			
Diseases	96	96	97
No diseases	99	98	99
f. Smoking			
Smoking	95	95	93
Non-smoking	96	92	98
Second Experiment			
Different demographics	97	98	97

non-smokers. For non-smokers, the accuracy of the model was 96%, and for smokers, the accuracy was 95%. The performance metrics of all similar groups for the diagnosis of COVID-19, as determined by the GA-SVM model, are presented in Table 2.

Figure 9 depicts an ROC curve for the GA-SVM model. When the age range was above 50 years, the AUC value was greater than 0.95 but less than 0.95 in the age range under 50 years. Regarding languages, the AUC value varied from 0.94 to 0.1. It was determined that the AUC value for symptomatic samples was 0.96, while for asymptomatic samples, it was 0.97. AUC values of 0.94 and 0.95 were obtained for the smoking and the non-smoking samples, respectively. Female and male groups both achieved an AUC value of 0.96. When there was a medical history of diseases, the AUC value was 0.97 versus 0.99 when there was no medical history of diseases.

5.2.2 Second Experiment

Based on the dataset, we conducted a second experiment to evaluate the GA-SVM model on various demographic characteristics of the samples. The model achieved an accuracy of 97%, a sensitivity of 98%, and a specificity of 97% as shown in Table 2. The ROC curve for this experiment, which is plotted in Figure 9, achieved an AUC value of 0.97.

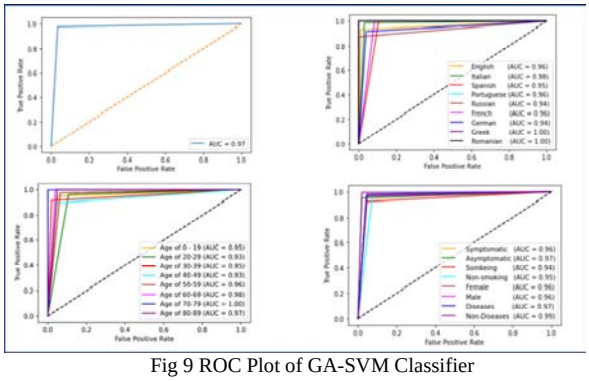

Fig 9 ROC Plot of GA-SVM Classifier

5.3 Discussion of Results

In this study, our principal findings and interpretations are based on the values of the analytical simulation during the experiments. The speech signals constitute an excellent predictor when screening for COVID-19 and have features that can be used as a diagnostic tool. This finding is consistent with other research findings regarding using speech as a diagnostic tool for COVID-19 [6] [7]. Additionally, based on the results, examining COVID-19 from speech using the same model in different languages and across all age groups is feasible.

The model detected the majority of COVID-19-positive patients, whether they were asymptomatic or experiencing symptoms. A further consideration is that the closer the ROC curve is to the upper left corner and when its value is greater than 0.9, the more efficient the model is for the test [43]. Figure 9 shows that the AUC values are greater than 0.9 in all of the ROC curves found in the experiments. Moreover, it has been observed that the K-means algorithm effectively reduces the number of dimensions and selects the best features from the feature space after the deep features based on Mel spectrogram images.

We have also discovered that speech utterances in sentences are more consistent and accurate predictors of COVID-19 infection than vowels, since vowels were also used in past studies to diagnose COVID-19 [19] [22]. As a consequence, the patients may find it difficult to produce speech, which may result in speech patterns or features that differ from those of a normal person.

Table 3 shows a comparison of the accuracy of the proposed methodology with the accuracy rates reported in

previous literature [19][20][21][22]. In contrast to the literature, our model obtained the audio features by extracting deep features from the Mel spectrogram images. In this case, deep features outperformed handcrafted or traditional features. We also compare some of the advantages and disadvantages of our features with those of the audio features proposed in previous literature, as presented in Table 3.

Based on our proposed methodology, we also performed a comparative analysis to highlight the diagnostic power of audio features in comparison to a previous study [21], which utilized the same data but added symptomatic features to the audio features for COVID-19 diagnosis. As a result, we integrated audio features and symptoms. To convert these symptoms into feature vectors, we indicated that individuals had symptoms (1) or (0) no symptoms.

The experimental results show that the accuracy of the acquired rate has decreased by 2% since our second experiment and achieved an accuracy of 95%. One possible explanation is that the symptoms do not correspond to the COVID-19 diagnosis. Figure 2 shows the data demographics, indicating 68% of the COVID-19-positive samples as asymptomatic, whereas 47% of the negative samples as symptomatic. We demonstrate that human speech contains hidden features that can be used to diagnose people who are asymptomatic or are experiencing symptoms.

Table 3 Comparison of Features Used in Our Proposed Methodology and Previous Studies

Model	Performance Metrics	Features	Advantages	Disadvantages
GA-SVM	Accuracy 97%	Deep features	Complex and robust features based on data and deep learning model. its high performance	Need more data
[19]	Accuracy 85%	Handcraf-ted features	Performance depends on the selected features and is suitable for another data type	Time-consuming
[20]	Accuracy 73%			
[21]	AUC value 0.79			
[22]	Accuracy 80%			

6. Conclusion

The rapid spread of COVID-19 and its high infection rates have overburdened global healthcare systems, aggravated by the high costs of clinical tests for COVID-19 and the long time it takes to obtain the results. Therefore, diagnosing COVID-19 by a cost-effective, fast, easy, and accurate method is crucial. Thus, the AI-based preliminary diagnosis for COVID-19 is regarded as a viable solution for taking the necessary preventive measures. Using audio samples of a person's speech in this study, we suggest a different AI-based diagnostic method for controlling the pandemic before infection transmission occurs among people.

In this work, we have presented proof of automatic detection of COVID-19 from human speech through a transfer learning technique to extract deep features from the audio dataset. The proposed methodology combines the pre-trained (VGG-16) model with Mel spectrogram images to extract deep features of the speech signals and the K-means algorithm that determines effective features. In addition to the GA-SVM classifier, it contains the GA for selecting the optimum hyperparameters for the SVM algorithm for diagnosing cases.

The proposed methodology performance was evaluated using a variety of evaluation criteria. The results show that the proposed techniques are suitable when the model is tested in different situations using the crowdsourced dataset. As a result, speech-based diagnosis can be one of the safest methods for COVID-19 diagnosis, which can help control the spread of the global pandemic.

In future work, we will evaluate the proposed model on other available COVID-19 audio datasets. We also plan to expand the datasets by diagnosing other diseases in addition to COVID-19.

Acknowledgments

The researchers would like to express their gratitude and appreciation to the Department of Computer Science and Technology at the University of Cambridge for making the dataset available to them in order for them to complete their research.

References

[1] D. Cucinotta and M. Vanelli, "WHO declares COVID-19 a pandemic," *Acta Biomed.*, vol. 91, no. 1, pp. 157–160, 2020, doi: 10.23750/abm.v91i1.9397.

[2] V. Corman *et al.*, "Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR," *Euro Surveill*, vol. 25, no. 3, pp. 1–8, 2020.

[3] R. W. Peeling, P. L. Olliaro, D. I. Boeras, and N.

Fongwen, “Scaling up COVID-19 rapid antigen tests: promises and challenges,” *Lancet Infect. Dis.*, vol. 21, no. 9, pp. e290–e295, 2021, doi: 10.1016/S1473-3099(21)00048-7.

[4] S. Swayamsiddha, K. Prashant, D. Shaw, and C. Mohanty, “The prospective of Artificial Intelligence in COVID-19 Pandemic,” *Health Technol. (Berl).*, vol. 11, no. 6, pp. 1311–1320, 2021, doi: 10.1007/s12553-021-00601-2.

[5] J. Shuja, E. Alanazi, W. Alasmary, and A. Alashaikh, “COVID-19 open source data sets : a comprehensive survey,” pp. 1296–1325, 2021.

[6] T. F. Quatieri, T. Talkar, and J. S. Palmer, “A Framework for Biomarkers of COVID-19 Based on Coordination of Speech-Production Subsystems,” *IEEE Open J. Eng. Med. Biol.*, vol. 1, pp. 203–206, 2020, doi: 10.1109/ojemb.2020.2998051.

[7] S. Sondhi *et al.*, “Voice Processing For Covid-19 Scanning And Prognostic Indicator,” *Heliyon*, vol. 7, no. 10, p. e08134, 2021, doi: 10.1016/j.heliyon.2021.e08134.

[8] Z. Ren, N. Cummins, J. Han, S. Schnieder, and J. Krajewski, “Evaluation of the Pain Level from Speech : Introducing a Novel Pain Database and Benchmarks Evaluation of the Pain Level from Speech : Introducing a Novel Pain Database and Benchmarks 1 Introduction,” no. October, 2018.

[9] L. Lin, X. Chen, Y. Shen, and L. Zhang, “Towards Automatic Depression Detection : A BiLSTM / 1D CNN-Based Model,” pp. 1–20, 2020, doi: 10.3390/app10238701.

[10] L. Verde, G. D. E. Pietro, and G. Sannino, “Voice Disorder Identification by Using Machine Learning Techniques,” *IEEE Access*, vol. 6, pp. 16246–16255, 2018, doi: 10.1109/ACCESS.2018.2816338.

[11] P. Vizza, G. Tradigo, D. Mirarchi, R. B. Bossio, and P. Veltri, “On the use of voice signals for studying sclerosis disease,” *Computers*, vol. 6, no. 4, pp. 1–12, 2017, doi: 10.3390/computers6040030.

[12] L. Zahid *et al.*, “A Spectrogram-Based Deep Feature Assisted Computer-Aided Diagnostic System for Parkinson’s Disease,” *IEEE Access*, vol. 8, pp. 35482–35495, 2020, doi: 10.1109/ACCESS.2020.2974008.

[13] M. Kiran Reddy *et al.*, “The automatic detection of heart failure using speech signals,” *Comput. Speech Lang.*, vol. 69, p. 101205, 2021, doi: 10.1016/j.csl.2021.101205.

[14] B. W. Schuller, D. M. Schuller, K. Qian, J. Liu, H. Zheng, and X. Li, “COVID-19 and Computer Audition: An Overview on What Speech & Sound Analysis Could Contribute in the SARS-CoV-2 Corona Crisis,” *Front. Digit. Heal.*, vol. 3, no. March, 2021, doi: 10.3389/fdgth.2021.564906.

[15] A. Imran *et al.*, “AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app,” *Informatics Med. Unlocked*, vol. 20, p. 100378, 2020, doi: 10.1016/j.imu.2020.100378.

[16] C. Brown *et al.*, “Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data,” *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 3474–3484, 2020, doi: 10.1145/3394486.3412865.

[17] H. Coppock, A. Gaskell, P. Tzirakis, A. Baird, L. Jones, and B. Schuller, “End- - to- - end convolutional neural network enables COVID-19 detection from breath and cough audio : a pilot study,” pp. 356–362, 2021, doi: 10.1136/bmjinnov-2021-000668.

[18] J. Han *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-Octob, no. October, pp. 4946–4950, 2020, doi: 10.21437/Interspeech.2020-2223.

[19] L. Verde, G. De Pietro, and G. Sannino, “Artificial Intelligence Techniques for the Non-invasive Detection of COVID-19 Through the Analysis of Voice Signals,” *Arab. J. Sci. Eng.*, vol. 19, 2021, doi: 10.1007/s13369-021-06041-4.

[20] M. Usman, V. K. Gunjan, M. Wajid, M. Zubair, and K. N. Siddiquee, “Speech as a Biomarker for COVID-19 Detection Using Machine Learning,” vol. 2022, 2022.

[21] J. Han *et al.*, “Exploring automatic covid-19 diagnosis via voice and symptoms from crowdsourced data,” *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2021-June, pp. 8328–8332, 2021, doi: 10.1109/ICASSP39728.2021.9414576.

[22] B. Stasak, Z. Huang, S. Razavi, D. Joachim, and J. Epps, “Automatic Detection of COVID-19 Based on Short-Duration Acoustic Smartphone Speech Analysis,” *J. Healthc. Informatics Res.*, vol. 5, no. 2, pp. 201–217, 2021, doi: 10.1007/s41666-020-00090-4.

[23] M. T. García-Ordás, J. A. Benítez-Andrades, I. García-Rodríguez, C. Benavides, and H. Alaiz-Moretón, “Detecting respiratory pathologies using convolutional neural networks and variational autoencoders for unbalancing data,” *Sensors (Switzerland)*, vol. 20, no. 4, 2020, doi: 10.3390/s20041214.

[24] L. Vavrek, M. Hires, D. Kumar, and P. Drotar, “Deep convolutional neural network for detection of pathological speech,” *SAMI 2021 - IEEE 19th World Symp. Appl. Mach. Intell. Informatics, Proc.*, pp. 245–249, 2021, doi: 10.1109/SAMI50585.2021.9378656.

[25] Q. Zhou *et al.*, “Cough Recognition Based on Mel-Spectrogram and Convolutional Neural Network,” *Front. Robot. AI*, vol. 8, no. May, pp. 1–7, 2021, doi: 10.3389/frobt.2021.580080.

[26] T. Xia *et al.*, “COVID-19 Sounds: A Large-Scale Audio Dataset for Digital Respiratory Screening,” *Thirty-fifth Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track (Round 2)*, pp. 1–13, 2021, [Online]. Available: https://covid19.who.int/.

[27] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2009.

[28] N. K. Manaswi, *Deep Learning with Applications Using Python Chatbots and Face, Object, and Speech Recognition With TensorFlow and Keras*. 2018.

[29] A. Solanki and S. Pandey, “Music instrument recognition using deep convolutional neural networks,” *Int. J. Inf. Technol.*, 2019, doi: 10.1007/s41870-019-00285-y.

[30] B. McFee *et al.*, “librosa: Audio and Music Signal Analysis in Python,” *Proc. 14th Python Sci. Conf.*, no.

Scipy, pp. 18–24, 2015, doi: 10.25080/majora-7b98e3ed-003.
[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.*, pp. 1–14, 2015.
[32] M. Elgendy, *Deep Learning for Vision Systems*. 2020.
[33] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. 2019.
[34] I. H. Witten and E. Frank, *Data Mining Practical Machine Learning Tools and Techniques*. 2005.
[35] H. Luu, *Machine Learning with Spark*. 2021.
[36] S. Kaur and S. Kalra, "Disease prediction using hybrid K-means and support vector machine," *India Int. Conf. Inf. Process. IICIP 2016 - Proc.*, 2017, doi: 10.1109/IICIP.2016.7975367.
[37] B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction by hybrid of K-means and extreme learning machine algorithms," *Expert Syst. Appl.*, vol. 11, no. 7, pp. 4581–4586, 2014.
[38] M. A. Khan, "An automated and fast system to identify COVID-19 from X-ray radiograph of the chest using image processing and machine learning," *Int. J. Imaging Syst. Technol.*, vol. 31, no. 2, pp. 499–508, 2021, doi: 10.1002/ima.22564.
[39] C. Cortes and V. Vapnik, "Support-Vector Networks," vol. 297, pp. 273–297, 1995.
[40] T. Agrawal, *Hyperparameter Optimization in Machine Learning: Make Your Machine Learning and Deep Learning Models More Efficient*. 2021.
[41] A. U. Haq *et al.*, "Feature Selection Based on L1-Norm Support Vector Machine and Effective Recognition System for Parkinson's Disease Using Voice Recordings," *IEEE Access*, vol. 7, no. March, pp. 37718–37734, 2019, doi: 10.1109/ACCESS.2019.2906350.
[42] E. Wirsansky, *Hands-On Genetic Algorithms with Python*. 2020.
[43] J. N. Mandrekar, "Receiver operating characteristic curve in diagnostic test assessment," *J. Thorac. Oncol.*, vol. 5, no. 9, pp. 1315–1316, 2010, doi: 10.1097/JTO.0b013e3181ec173d.

Aseel Alfaidi received a BS degree in computer science from Taibah University, Medina, Saudi Arabia, in 2018. She is currently pursuing an MS degree in computer science at the Department of Computer Science and the Artificial Intelligence University of Jeddah, Jeddah, Saudi Arabia. Her research interest includes Artificial Intelligence in healthcare and voice analysis for disease diagnosis.

Dr. Abdullah Alshahrani holds his BS degree in Computer Science from King Khalid University, in 2007, and received an MS degree of Computer Science from School of Engineering Mathematical Sciences, La Trobe University, Australia, in 2010, and his Ph.D. in Computer Science from The Catholic University of America, Washington DC, in 2018. He is currently an assistant professor in the Department of Computer Science and Artificial Intelligence, College of Computer Science and Engineering, University of Jeddah, Jeddah 21493, Saudi Arabia.

Dr. Maha Aljohani holds her BS degree in Information Systems from Taibah University, in 2009 and received an M.S. degree in Computer Science from Dalhousie University, Canada in 2013, and her Ph.D. in Computer Science from Dalhousie University, Canada in 2019. She is currently an assistant professor in the Department of Software Engineering, College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia.