Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Journal of Biomedical Informatics 130 (2022) 104078

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Diagnosis of COVID-19 via acoustic analysis and artificial intelligence by monitoring breath sounds on smartphones

Zhiang Chen, Muyun Li, Ruoyu Wang, Wenzhuo Sun, Jiayi Liu, Haiyang Li, Tianxin Wang, Yuan Lian, Jiaqian Zhang, Xinheng Wang *

School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China

ARTICLE INFO

Keywords
COVID-19
Breath sound
Acoustic analysis
Convolutional Neural Network (CNN)
k-Nearest Neighbors (*k*NN)

ABSTRACT

Scientific evidence shows that acoustic analysis could be an indicator for diagnosing COVID-19. From analyzing recorded breath sounds on smartphones, it is discovered that patients with COVID-19 have different patterns in both the time domain and frequency domain. These patterns are used in this paper to diagnose the infection of COVID-19. Statistics of the sound signals, analysis in the frequency domain, and Mel-Frequency Cepstral Coefficients (MFCCs) are then calculated and applied in two classifiers, *k*-Nearest Neighbors (*k*NN) and Convolutional Neural Network (CNN), to diagnose whether a user is contracted with COVID-19 or not. Test results show that, amazingly, an accuracy of over 97% could be achieved with a CNN classifier and more than 85% on *k*NN with optimized features. Optimization methods for selecting the best features and using various metrics to evaluate the performance are also demonstrated in this paper. Owing to the high accuracy of the CNN model, the CNN model was implemented in an Android app to diagnose COVID-19 with a probability to indicate the confidence level. The initial medical test shows a similar test result between the method proposed in this paper and the lateral flow method, which indicates that the proposed method is feasible and effective. Because of the use of breath sound and tested on the smartphone, this method could be used by everybody regardless of the availability of other medical resources, which could be a powerful tool for society to diagnose COVID-19.

1. Introduction

Fighting against COVID-19 is still the most urgent issue across the world because of its serious damage to human health and even life. In order to fight against COVID-19, the first important thing is to diagnose the patient who has contracted COVID-19 and then treat them according to the conditions of the infection.

At present, the main diagnostic methods are nucleic acid test, antibody test, and antigen test [1]. These methods are effective and accurate. Particularly the nucleic acid test is the"gold standard" of diagnosis [1]. However, these diagnosis methods have application shortcomings. The first problem is its long testing process owing to the collection of samples, sending samples to test labs with professional equipment, and release of the result. This process needs at least a few hours, sometimes a few days, even up to 10 days in some countries, as laboratories become overwhelmed [2,3]. During this process, patients with COVID-19 could become spreaders. The second one is that people have to travel to a sample collection point, which makes them more likely to be exposed to coronavirus. The third one is that these methods are not good for people living remotely to use because of the lack of medical professionals to provide advice and help in collecting samples. New diagnosis methods are desperately needed.

Recently some other methods were developed based on other physical phenomena, for example, based on quality of voice when a sustained vowel/a/ was pronounced and analyzed [4], and a series of vowels/i:/ ,/e:/,/o:/,/u:/, and/a:/ was recorded and analyzed [5]. Similar to voice, speech is also an approach to diagnosing COVID-19. The latest research showed that 88.2% accuracy could be achieved on voice [6], and 82–86% accuracy on speech [7].

Apart from voice and speech, the sound is also used to diagnose COVID-19. Pioneering research in using quality of sound to diagnose the infection of COVID-19 is to analye the cough sounds by scientists at MIT [8]. This diagnosis approach was verified from other studies [1,9–12].

One important tool behind these diagnosis methods is artificial intelligence (AI). The power of AI in classification made these approaches feasible. Apart from diagnosis, AI is also used in Covid-19 research for

* Corresponding author.
E-mail address: xinheng.wang@xjtlu.edu.cn (X. Wang).

https://doi.org/10.1016/j.jbi.2022.104078
Received 24 November 2021; Received in revised form 9 April 2022; Accepted 16 April 2022
Available online 27 April 2022

outbreak detection and biomarker discovery [13]. This is confirmed by the latest research results, where researchers from Stellenbosch University [14] and Imperial College London [15] have successfully applied deep transfer learning and end-to-end convolutional neural network to the study of COVID-19 detection, respectively.

Diagnosis from voice, sound, and coughing provides an approach that could be used by a vast majority, particularly for those living remotely because of no need for professional equipment and collection of samples. However, diagnosis based on the voice needs specific pronunciations, so the tester should be trained before. Diagnosis based on speech [16] has a similar working principle to diagnosis based on the voice that is investigating short-duration speech segments (e.g., held vowel, nasal phrase), but more than 90% test accuracy is not achievable. Coughing is not a symptom of all COVID-19 patients, manifesting itself only in 67.7% of cases [17]. The latest discovery from Omicron variants in the UK also found that the top symptoms are runny nose, headache, fatigue, sneezing, and sore throat. Only 44% of people reported a persistent cough [18]. In contrast, many patients with the disease have obvious early signs of lung pathology before the onset of symptoms such as dry cough, fever, and dyspnea [19]. These lung lesion symptoms include a peripheral distribution (80%), ground-glass opacity (91%), and vascular thickening (59%), which will produce changes in the respiratory sound [20]. Therefore, in this paper, we have proposed another approach for diagnosing COVID-19, which is based on breathing. We believe diagnosis based on breathing is more appropriate because breathing is natural. It doesn't need any training for users.

From analyzing breathing signals and training with two algorithms, it verified that this approach is feasible. With the optimized training model, the diagnosis could reach more than 97%. In addition, we have implemented the diagnosis method on smartphones, which could be used by anybody with a smartphone, anywhere and anytime. The initial medical test indicates that the result from breathing is quite similar to lateral flow. This will be particularly useful for mass diagnosis or for those living remotely.

In a summary, this paper has the following contributions: (1) Acoustic analysis of breath sound from COVID-19 patients were conducted and change of frequency components was discovered, which provides a fundamental evidence for applying acoustic analysis for diagnosing COVID-19; (2) Deep learning classifier, convolutional neural network (CNN) in this case, provides a promising diagnosis tool, where accuracy reaches over 97% with the right features from acoustic analysis; (3) MFCC is confirmed an effective feature for diagnosing COVID-19 from breath sound; (4) The designed CNN model needs to be trained with optimizations by considering data type, imbalance of data, features, pre-processing of data, and length of data to obtain a best performance model; (5) The trained model could be implemented on smartphones to provide a fast turnaround test result, within two minutes even with a low performance smartphone; (6) Initial medical test verified the effectiveness of this diagnosis method by comparing with biological method.

The remaining of this paper is organized as follows: following the introduction, technical details of the diagnosis method will be described in Section 2 Methods. Test results and performance evaluation, including test results after implementing the CNN model on smartphones, will be presented in Section 3 Results. Further results analysis will be presented in Section 4 Discussion. Finally, this paper concludes in Section 5 Conclusions.

2. Methods

Technical details, including the dataset of breath sound used for the research, signal analysis of the breath sound of both healthy people and patients with COVID-19, diagnosis method with two artificial intelligence algorithms, *k*-Nearest Neighbors (*k*NN) and CNN are presented in this section.

2.1. Dataset

A dataset called Coswara-Data from the Indian Institute of Science (IISc) Bangalore was used for signal analysis, training and testing [21]. As a publicly available dataset, it asked participants to provide recordings of fast and slow breathing sounds, deep and shallow coughing sounds, sustained phonation of vowels, and counting exercises at a slow and a fast pace, and each data was labeled with health status and other clinical information.

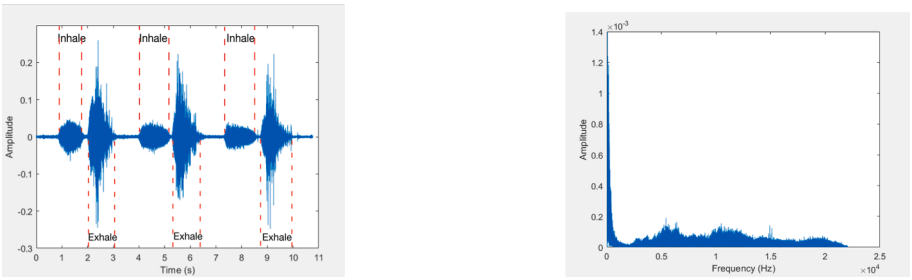
This dataset includes 1107 healthy providers, 107 COVID-19 positive patients, 224 providers with uncertain health status, and 48 patients with other respiratory diseases but negative COVID-19 tests. Other information in the dataset is also worth considering in order to avoid bias in model training. In terms of gender distribution, there were 1123 males and 363 females among the data providers. On the age distribution, the providers were mainly concentrated in the age range of 18 to 30 years old, and decreasing in order. The oldest reaches 70 to 80 years old, and the youngest includes a portion of minors younger than 18 years old. In terms of geographical distribution, most of the data come from six states of India, including Karnataka, Maharashtra, Tamil Nadu, West Bengal, Telangana, and Kerala.

To complement the disease identification method and to improve disease identification in different situations, for the choice of audio type, deep breathing audios are chosen in this research. The sound of respiratory activity can be effectively used to assess the health of the lungs, and deep breathing can better expose the respiratory characteristics of the subject and facilitate medical diagnosis as opposed to ordinary breathing sounds [10]. Deep breathing data were collected at a sampling rate of 48 kHz and a resolution of 16 bits.

2.2. Signal Analysis of Breath Sounds

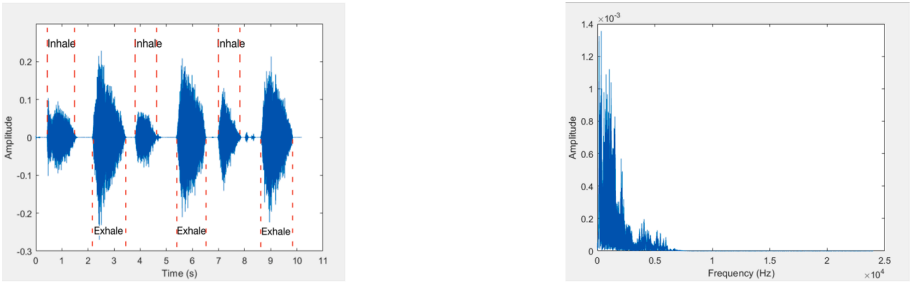
Breath sounds from 10 healthy people and 10 people with COVID-19 were examined manually in the time domain and frequency domain to analyze the pattern. Fig. 1(a) shows a typical healthy breath sound signal and its corresponding spectrum derived from Fast Fourier Transform (FFT) is shown in Fig. 1(b). Figs. 1(b) and 2(b) show one typical type of breath sound with COVID-19 but show no symptom and its spectrum; Fig. 2(b) and Fig. 3(b) show another type of COVID-19 signal with symptom and its spectrum, respectively. From direct observation, it can be found that the signals with COVID-19 differ significantly from the healthy ones. For the healthy breath sound, the strength of inhaling is very low, compared to exhaling. However, for the breath sounds with COVID-19, the strength of inhaling increases even if there is no symptom, and the strengths of exhaling and inhaling are almost the same for signals with COVID-19 symptoms, which means patients need to breath harder owing to the malfunction of the lungs.

From the spectrum of healthy breath sound and sound with COVID-19, it can be seen that the main frequency components of the healthy breath sound are distributed between 50 Hz and 5000 Hz, and the low-frequency components have higher amplitudes, as shown in Fig. 1(b). However, sounds with COVID-19 generally have two cases of respiratory sounds. The first case has a similar frequency distribution to healthy breathing sound, but with more high energy components appearing at a higher frequency, even outside the normal healthy breath sound range, as shown in Fig. 2(b). The second case is that the high-frequency part of the breathing sound has concentrated high energy, and the low-frequency part has very little energy, as shown in Fig. 3(b). Through the analysis of more than 30 samples, we found that the phenomenon of high-frequency energy concentration in the unhealthy breathing sound is caused by the symptoms of breathing difficulties and nasal congestion, which result in the elevation of breath sounds. The condition similar to healthy breath sound is because the patient has no clinical symptoms such as fever and dyspnea, which is an asymptomatic infection. These results verified the clinical evidence that acoustics of breath sound had changed after infecting COVID-19 and acoustics could be used for the



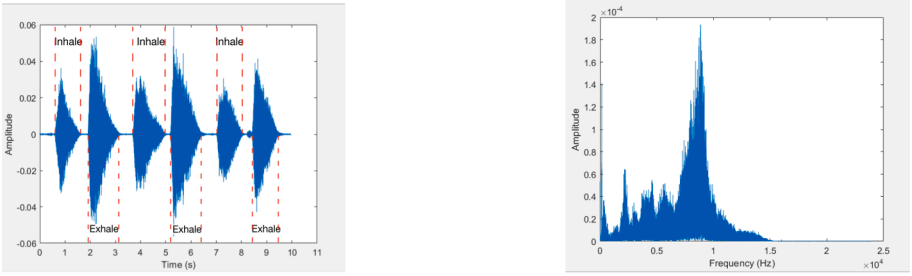
(a) (b)

Fig. 1. (a) Healthy breath sound in time domain (b) Spectrum of healthy breath sound.



(a) (b)

Fig. 2. (a) Breath sound with COVID-19 in time domain (no symptom) (b) Spectrum of breath sound with COVID-19.



(a) (b)

Fig. 3. (a) Breath sound with COVID-19 in time domain (showing symptom) (b) Spectrum of breath sound with COVID-19.

diagnosis of COVID-19 [4].

2.3. Diagnosis Method of COVID-19

In order to diagnose and classify COVID-19, a process was designed with four steps. After recording breath sound on a smartphone, the first step is to pre-process the raw data. The second step is to extract features from the pre-processed data. The third step is to use the features to train the classifiers, and the last step is to verify the testing result. The schematic diagram is illustrated in Fig. 4. Details of each step will be described in each sub-section.

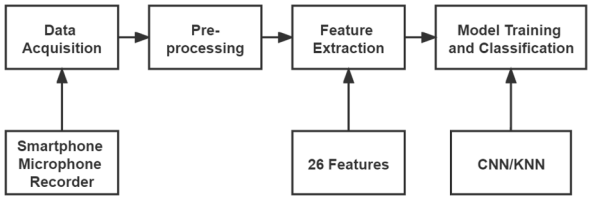


Fig. 4. Classification Flow Chart.

2.3.1. Signal Pre-processing

When the breath sound is recorded, the analog signal is digitalized. Quantization noise and distortion will be brought in the quantization process of the digitalized sound signal. The purpose of pre-processing is to reduce the influence of aliasing and high-order harmonic distortion from recorded signals. Four key pre-processing techniques are applied in this paper, including pre-emphasis, normalization, framing and windowing, and noise reduction.

2.3.1.1. Pre-emphasis. Most of the radiative effects of breath sound come from the lips, which results in normally lower amplitudes in the high-frequency components. A first-order FIR high-pass digital filter [22] was implemented to increase the spectral energy of the high-frequency component, which is defined as

$$x(i) = x(i) - \alpha * x(i-1) \tag{1}$$

where $x(i)$ represents the raw data, i represents the $i-th$ sampling point, α is the pre-emphasis factor, usually $0.9 < \alpha < 1$. Here α is taken as 0.97. The transfer function of this filter is $H(z) = 1 - \alpha z^{-1}$, where z donates Z-transform ($z = e^{j\omega}$).

2.3.1.2. Normalization. The raw data collected from different devices have different specifications, which will affect feature extraction. Therefore, the signal needs to be normalized by following (2) to reduce the impact.

$$x_{norm}(i) = \frac{x(i)}{max(|x(i)|)} \tag{2}$$

where $x(i)$ represents the raw data, i represents the $i-th$ sampling point, $max(|x(i)|)$ represents the maximum absolute value of the raw data $x(i)$, and $x_{norm}(i)$ is the normalized data.

2.3.1.3. Framing and Windowing. The requirement of the Fourier transform is that the input signal has to be stationary. The time-varying breathing signals can be considered to be approximately constant in a short period of time (generally 10–30 ms), that is, breathing signals have short-term stability. Here the frame length to 32 ms and frameshift to 50% were adopted. At the same time, in order to reduce the leakage in the frequency domain after framing, a rectangle window function was introduced by following (3):

$$w(n) = \begin{cases} 1, & 0 \leqslant n \leqslant N-1 \\ 0, & otherwise \end{cases} \tag{3}$$

where N is the frame size, which can be calculated as $32m \times$ sampling rate$= 32m \times 48k = 1536, w(n)$ is the function of the rectangle window and n represents the sampling points in each frame.

2.3.1.4. Noise reduction. The acquired sound signal is the superposition of the original signal and the noise signal. These additive noise signals are either smooth or slowly changing and could be removed to extract the breath sound. The most commonly used method in speech denoising - spectral subtraction - was applied to denoise the recorded signals. As shown in Fig. 5(a) and (b), by comparing the spectrogram, it can be found that the color of the de-noised signal becomes lighter in the voiceless segment, which indicates that the noise energy is reduced.

2.3.2. Feature Extraction

Unlike speech signal, breath sound is a kind of bandwidth noise. Therefore, feature extraction is the most important part before classification. Common feature extraction methods include Mel-Frequency Cepstral Coefficients (MFCCs), wavelet transform, autoregressions modeling, etc. [23–25]. MFCCs are adopted as one group of main features. MFCCs are a set of features widely used in audio and speech processing [26], which are superior to the linear prediction coefficient (LPC). These features will be used in this paper to classify the breath sound signals. Apart from MFCCs, clear evidence shows that the shape of breath sound in both the time domain and frequency domain has been changed after the contraction of COVID-19. Because of the change of shapes, statistics of the signals and parameters used to differentiate the shape of the signals in both the time domain and frequency domain are also used as features in this paper to classify the signals, which include Mean Value, Standard Deviation, Mean Absolute Deviation, Quantile 25, Quantile 75, Interquartile Range, Skewness, Kurtosis, Signal Entropy, Spectral Entropy, Dominant Frequency Value, Dominant Frequency Magnitude, and Dominant Frequency Ratio. Because they are standard parameters, details for calculating them are not presented here.

MFCCs are based on the logarithmic spectrum expressed in a nonlinear Mel scale and its linear cosine transform. In this paper, the first 13 coefficients are used.

Given a pre-processed signal $x'(i)$, from a raw signal $x(i)$, after FFT transform, amplitude spectrum of the signal becomes:

$$X(n) = \sum_{n=0}^{N-1} x'(i) e^{-j2\pi n/N}, 0 \leqslant n \leqslant N \tag{4}$$

where $x'(i)$ is the pre-processed signal, N is the frame size and n represents the sampling points in each frame.

Power spectrum, P, is obtained as:

$$P(n) = \frac{1}{N}\left|X(n)^2\right| \tag{5}$$

where N is the frame size and n represents the sampling points in each frame.

After calculating the energy spectrum of each frame, the energy spectrum is passed through a Mel-scale triangular filter bank to smooth the spectrum, eliminate the effect of harmonics, and highlight the original speech resonance peak [26]. The specific process is to, firstly, convert the signal from the actual frequency to Mel-frequency, and then pass the power spectrum through 26 Mel-scale triangle filters, and the obtained value is the energy value of the frame data in the corresponding frequency band of the filter.

The Mel-frequency is defined as:

$$F_{mel}(f) = 1125ln\left(1 + \frac{f}{700}\right) \tag{6}$$

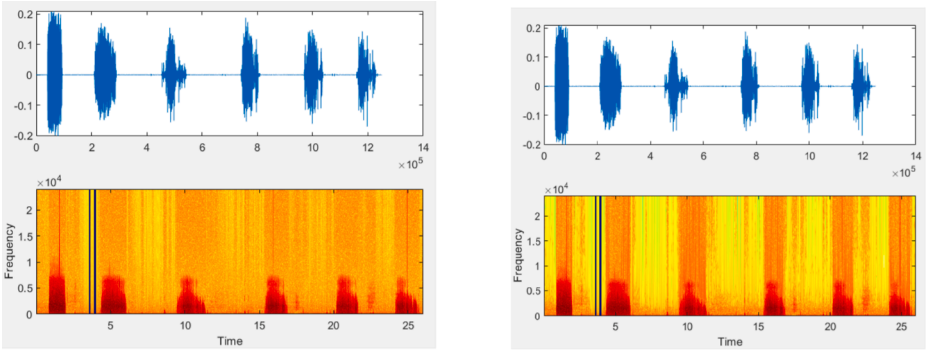
where F_{mel} is the perceived frequency in Mel, and f is the actual frequency in Hz.

The center frequency of the triangle filter is defined as:

$$f(m) = \left(\frac{N}{f_s}\right) F_{mel}^{-1}\left(F_{mel}(f_l) + m\frac{F_{mel}(f_h) - F_{mel}(f_l)}{M+1}\right) \tag{7}$$

where f_l and f_h are the lowest and the highest frequency of the filter, respectively, N is the length of FFT, M represents the total number of filters, which is 26.

The frequency response of the triangle filter is defined as:



(a) (b)

Fig. 5. (a) Original signal (b) De-noised signal.

$$H(n)=\begin{cases}0 & ,n<f(m-1)\\ \dfrac{2(n-f(m-1))}{(f(m+1)-f(m-1))(f(m)-f(m-1))} & ,f(m-1)\leqslant n\leqslant f(m)\\ \dfrac{2(f(m+1)-n)}{(f(m+1)-f(m-1))(f(m)-f(m-1))} & ,f(m\leqslant n\leqslant f(m+1))\\ 0 & ,n\geqslant f(m+1)\end{cases} \tag{8}$$

where $m=1,2,\ldots,26$ represents 26 triangle filters, n represents the sampling points in each frame, and $f(m)$ is the center frequency of the triangle filter.

Taking the logarithm of the energy output of each filter bank, it yields:

$$s\left(m\right)=ln\left(\sum_{n=0}^{N-1}|X(n)|^2H\left(n\right)\right),0\leqslant m\leqslant 26 \tag{9}$$

where $s(m)$ represents the logarithmic energy of the output of each filter bank, $m=1,2,\ldots,26$ represents 26 triangle filters, N is the frame size and n represents the sampling points in each frame.

The above logarithmic energy is put into the discrete cosine transform (DCT) to find the L-order Mel cepstrum coefficient:

$$C(k)=\sum_{m=0}^{N-1}s(m)\cos(\frac{\pi k(m-0.5)}{26}) \tag{10}$$

where $k=1,2,\ldots L$ represents the order of MFCC coefficient, usually setting to 12–16, and here we set $L=13$ to take 13 of 26 coefficients from 26 triangle filters and N is the frame size.

Fig. 5(a) and (b), and Fig. 6(b) show the graphic results obtained by using MFCC feature extraction. The vertical axis of the image represents 13 MFCCs, and the horizontal axis represents time, meaning the change of MFCCs in the time domain. By comparison, we can see that there are differences between healthy breath sound and breath sound with COVID-19.

2.3.3. Classification Models

Previous studies have shown that sound analysis is an effective method to detect and diagnose various respiratory diseases. In order to explore the advantages of breath sound diagnosis and resolve the problem based on subjective auscultation diagnosis, researchers have applied various automated signal processing and classification methods [23,27–32]. For example, in [27], two different machine learning techniques: Gaussian Mixture Model (GMM) and Support Vector Machine (SVM) were employed to extract acoustic features for speech modeling of an adolescent with depression. A similar system was proposed in [28], which combined MFCCs to extract and classify voice features, resulting in an accuracy of 96.1% in the diagnosis of voice diseases. In a recent study of COVID-19 detection in cough, breath and speech [14], three pre-trained deep neural networks: CNN, LSTM, and Resnet50 combined with deep transfer learning achieved the highest AUCs of 0.982.

In this paper, two classifiers were designed to determine whether a user has suffered from COVID-19 or not. One is *k*NN and the other is CNN. The reason to select *k*NN is to use a simple-to-implement algorithm in smartphones with low computing resources so that all the smartphone carriers may benefit from this diagnosis method.

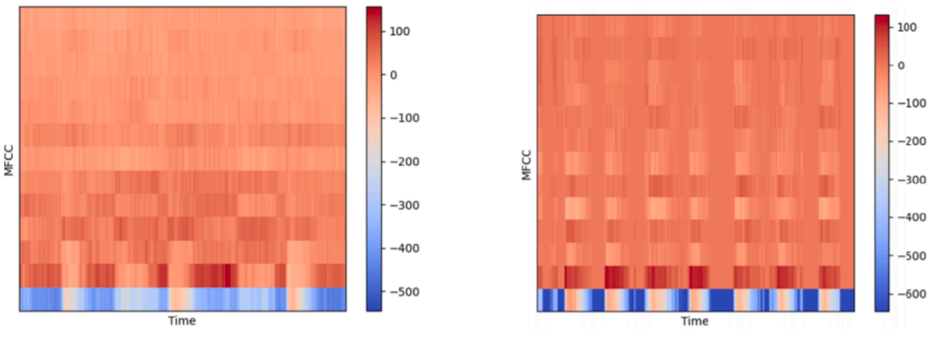
2.3.3.1. k-Nearest Neighbors (kNN). *k*NN is one of the simplest and most commonly used supervised learning classification algorithms in data mining. The idea of this algorithm is very intuitive: if most of the k closest samples in the feature space belong to a certain category, then the sample also belongs to this category. Due to the non-parametric and inert characteristics of *k*NN, the *k*NN algorithm used in this paper to build the model has the advantages of fast training time, easy to use, and good prediction effect.

After pre-processing, all pre-processed signals are divided into audio chunks of 3 s in length with labels, where 3-s chunks are supposed to be the best for the CNN model. These signals are then input into the *k*NN model. The validation method is the hold-out validation with a 25% hold-out percentage. This *k*NN model has a number of neighbors of 10, and a distance metric of Euclidean, which measures the absolute distance between points in a multidimensional space, and distance weight is equal.

Here three different *k*NN models are used for classification.

Model 1: 26 features extracted from the audio are used for classification, without any other optimization techniques.

Model 2: In order to improve the performance, the first approach adopted was to find out the effective features used for training this model. To this end, a parallel coordinate plot was implemented to evaluate all 26 features on four metrics, True Negative, True Positive, False Negative, and False Positive. Each feature is scaled to a different range in a numerical form, which represents the feature value. By comparing all 26 feature values, it can be initially determined that the Features Standard Deviation, Mean Absolute Deviation, Quantile75, Signal IQR, Dominant Frequency Magnitude, and Dominant Frequency Ratio, are likely to have no positive effect on the classification because their feature values are distributed in a uniformly dispersed state without particularity. Therefore, we can conclude that not all 26 features are valid when training the model. However, the large number of feature values produced by this approach is inconvenient to observe and is crude. In order to further optimize this model, the Neighbourhood Component Analysis (NCA) method was further used to analyze the data, where the importance of all 26 features is shown in Fig. 7. The horizontal coordinates represent each feature in sorted order and the vertical coordinates represent the importance of this feature. By making this selection, features were reduced from 26 to 14. Selected features include Sample Kurtosis (horizontal coordinate 8), Signal Entropy



(a) (b)

Fig. 6. (a) MFCCs of healthy breath sound (b) MFCCs of sound with COVID-19.

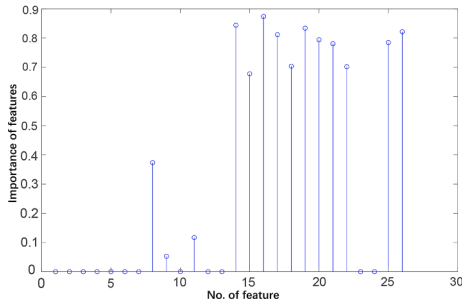


Fig. 7. Feature Selection.

(horizontal coordinate 9), Dominant Frequency Values (horizontal coordinate 11), MFCC1 to MFCC9 (horizontal coordinate 14 to 22), MFCC12 (horizontal coordinate 25), and MFCC13 (horizontal coordinate 26).

The selected 14 features were again put into the *k*NN model and tuned using Bayesian optimization with the number of iterations setting to 45. As can be seen in Fig. 8, the Best point hyperparameters and the Minimum error hyperparameters appear simultaneously at the 20*th* iteration. In this model, the number of neighbors is 1, the distance metric is City block, and the distance weight is Inverse. This optimized model is noted as Model 2.

Model 3: The performance of Model 2 has been improved. However, considering that the model will be used for the diagnosis of disease, the false-negative rate of the confusion matrix requires additional attention. This is because the consequences of a positive patient being incorrectly classified as negative are more serious than the consequences of a healthy person being incorrectly classified as positive in practice. To address this problem in a targeted manner, a misclassification cost matrix was introduced in this model. The expected cost $L(a,i)$ of sample a being classified as class i can be expressed as:

$$L(a,i) = \sum_{j} P(j|a)C(i,j) \tag{11}$$

where a is an sample, (a,i) denotes its classification into category i, $P(j|a)$ denotes the posterior probability obtained in the algorithm that a belongs to category j, and $C(i,j)$ denotes the real cost of the algorithm misclassifying a sample from category i as category j.

In such an equation, because of the inclusion of the loss weight bias factor $C(i,j)$, the objective model will not only simply focus on how to obtain the maximum value of $P(j|a)$, but also take into account both the predicted outcome $P(j|a)$ and the loss $C(i,j)$ caused by the different predicted outcomes $C(i,j)$, resulting in these two factors holding each other in check.

2.3.3.2. Convolutional Neural Network (CNN). CNN, as an effective tool for the analysis of visual imagery [33], was applied in this paper to

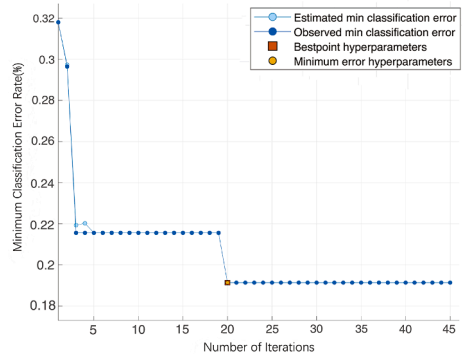


Fig. 8. Minimum Classification Error Plot.

classify the differences between MFCCs from healthy and unhealthy people. After that, it will be able to give pre-screening diagnostics. Our proposed model, as shown in Fig. 9, takes a recording of breaths, performs signal pre-processing, and inputs features into a CNN model to output a pre-screening diagnostic.

At the pre-processing stage, each recorded breath sound is split into 3-s audio chunks, padded as needed, processed with the MFCC package [34] and subsequently passed into a CNN as described in the next paragraph.

The CNN architecture is mainly made up of one ResNet50; there are 2048 classes for the ResNet50, which can be interpreted as one-dimensional features extracted from MFCCs. These features are then fed into a 1024 neuron deeply connected neural network layer (dense) with a ReLU activation function, and finally a binary dense layer with a sigmoid activation function.

For our experiments, the CNN model was trained in a computing environment with eight CPU cores and one GPU core. The model of the CPU is Intel ® Xeon ® Platinum 8255C CPU @ 2.50 GHz. The L1 cache is 32 KB + 32 KB; L2 cache is 4096 KB; L3 cache is 36608 KB. For each CPU core, 4.5 GB memory is assigned. That is to say, we have 36 GB memory available with a total of eight CPU cores. The model of the GPU is NVIDIA ® Tesla ® V100 SXM2 32 GB Computational Accelerator. It has a memory of 32 GB HBM2 and the memory bandwidth is 900 GB/s.

To effectively examine the feasibility of the proposed method, two techniques were adopted when the CNN model was trained.

Firstly, since the size of the dataset is relatively small, data augmentation was performed. Unlike the conventional CNN task where the photos may be rotated or re-scaled to increase the number of training samples, this work requires a different data augmentation method because the input features are MFCCs, which become meaningless when processed in these ways. Therefore, when the breath sound was split into 3-s audio chunks, a 90% overlapping was employed. The size of our training dataset was expanded almost 10 times with this data augmentation, which is a significant help in training the model and improving accuracy.

After the training data was augmented, it was found that the dataset was slightly imbalanced, as shown in Table 1. In order to optimize the training effectiveness and the model performance, Mean Squared False Error (MSFE) was chosen as the loss function when training the best-optimized model [35]. According to research in [35], MSFE outperforms the conventional Mean Squared Error (MSE) at a class-imbalanced problem. Our experiment supports this conclusion.

CNN is a very powerful tool for classifying objects. CNN is employed in this paper as another classifier to evaluate the effectiveness of diagnosing COVID-19. By using CNN, questions to be answered include: (1) whether recorded raw breath sound could be implemented directly on CNN or not? (2) whether pre-processing of raw data will improve the performance or remove the features or not? and (3) whether other features representing different properties of the signals make the classification more accurate and efficient or not?.

To answer the question (1), two approaches were adopted in this paper to input the data into the CNN model; one is to input the raw data directly, and the other is to extract the MFCCs and then use MFCCs as input to the CNN model.

To answer the question (2), signal pre-processing needs to be evaluated. Considering that pre-processing might exclude features that contribute to the classification between unhealthy and healthy users, we want to first verify whether pre-emphasis and normalization can increase the prediction accuracy or not. And after evaluating the necessity of including pre-emphasis and normalization, we want to explore whether noise reduction can help improve the model performance or not. We take the previously analyzed model with both pre-emphasis and normalization as a baseline.

To answer the question (3), several variants of MFCC were applied for classification.

To make it possible to diagnose COVID-19 on a smartphone through

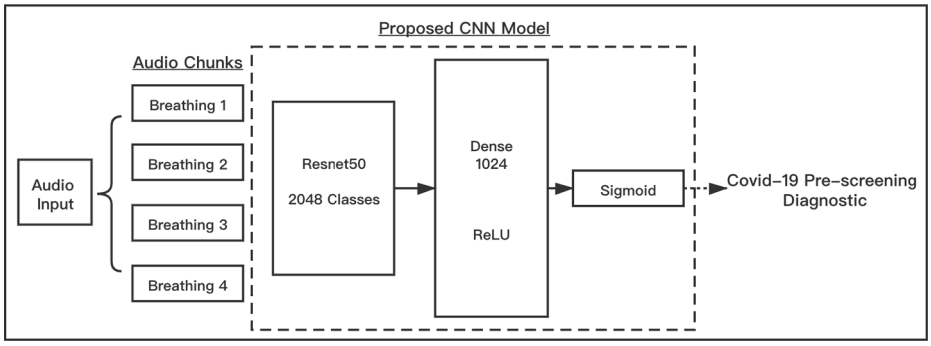


Fig. 9. Architecture of CNN Model.

Table 1
Imbalanced Training Data.

COVID-19 Status	Negative	Positive	Total
Number of Audio Chunks	3974	2774	6748
Percentage	58.9%	41.1%	100.0%

acoustic analysis, we have managed to develop an Android app incorporating the trained CNN model, which can be downloaded from [36]. The design of the app is very simple with only a button to start the recording, a button to stop the recording, and then a button to do the diagnosis, where the screenshot of the app is shown in Fig. 10. The possibility of infecting COVID-19 is presented. For calculating the probability, only three steps are required. Firstly, the input audio is split into 3-s audio chunks with 90% overlapping. Secondly, these audio chunks are used as the input for CNN inference. After the inference calculation of the neural network, the probability of each 3-s segment will be directly output. Finally, the average of all these inference results is calculated as below:

$$\mathrm{P} = \frac{1}{n}\sum_{i=1}^{n} I(x_i) \tag{12}$$

where n represents the total number of chunks that the user's breath sound can be split with 3-s length and 90% overlapping, x_i represents the $i-th$ input chunk of the user and $i = 1 \ldots n$. $I(x_i)$ represents the output of neural network, which refers to the probability that the input chunk x_i. P represents the probability that the user is infected with COVID-19 virus.

One example shown in Fig. 10 is the probability of 0.12, which is the test result from one of the authors and shows a very low possibility of infecting COVID-19. The calculation time depends on the types of smartphones and also the length of breath sounds recorded. It normally takes less than one minute to complete the calculation. If the length of recorded breaths is long, such as far more than three seconds, it will take a longer time to finish it but normally in less than 2 min on a reasonably powerful smartphone.

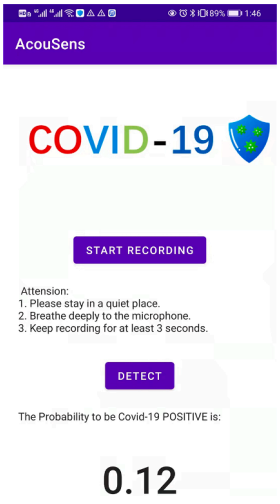


Fig. 10. Screenshot of the application.

3. Results

The complete diagnosis method was introduced in the previous section. In this section, results are going to be presented to evaluate the effectiveness and performance of the proposed method. Firstly the evaluation method is going to be introduced. Then the performance of the two models is evaluated. Finally, the implementation of CNN on Android phones and its use of it in the initial medical test are presented to verify the application of the proposed method.

3.1. Evaluation Method

In order to evaluate the performance of our proposed model, 75% of randomly selected audio chunks were used for training and the rest 25% were used for testing. The test accuracy, confusion matrix, Receiver Operating Characteristic (ROC), and Area under the ROC Curve (AUC) can then be obtained for quantitative assessment.

Test accuracy is defined from the four parameters, namely True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), of the confusion matrix, as:

$$\text{Test Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \tag{13}$$

In case of multiple times of training for one model are conducted, accuracy is determined from the average of each training.

3.2. Model Performance

3.2.1. kNN performance

Model 1: As seen from the table that the accuracy of *k*NN model 1 reached only 70.0%, while the True Positive Rate ($TPR = \frac{TP}{TP+FP} \times 100$) reached only 53.6%, which is not a satisfactory result.

Model 2: After optimisation of features, the performance is shown in Table 2 under label Model 2. It can be seen that an increase of 10.1% in accuracy was achieved. The Validation ROC Curve is shown in Fig. 11

Table 2
*k*NN Model Performance.

Model	Number of features selected	Accuracy	AUC	TPR	TNR
1	26	70.0%	0.78	53.6%	80.8%
2	14	80.1%	0.80	71.2%	85.9%
3	14	78.0%	0.83	82.4%	75.2%

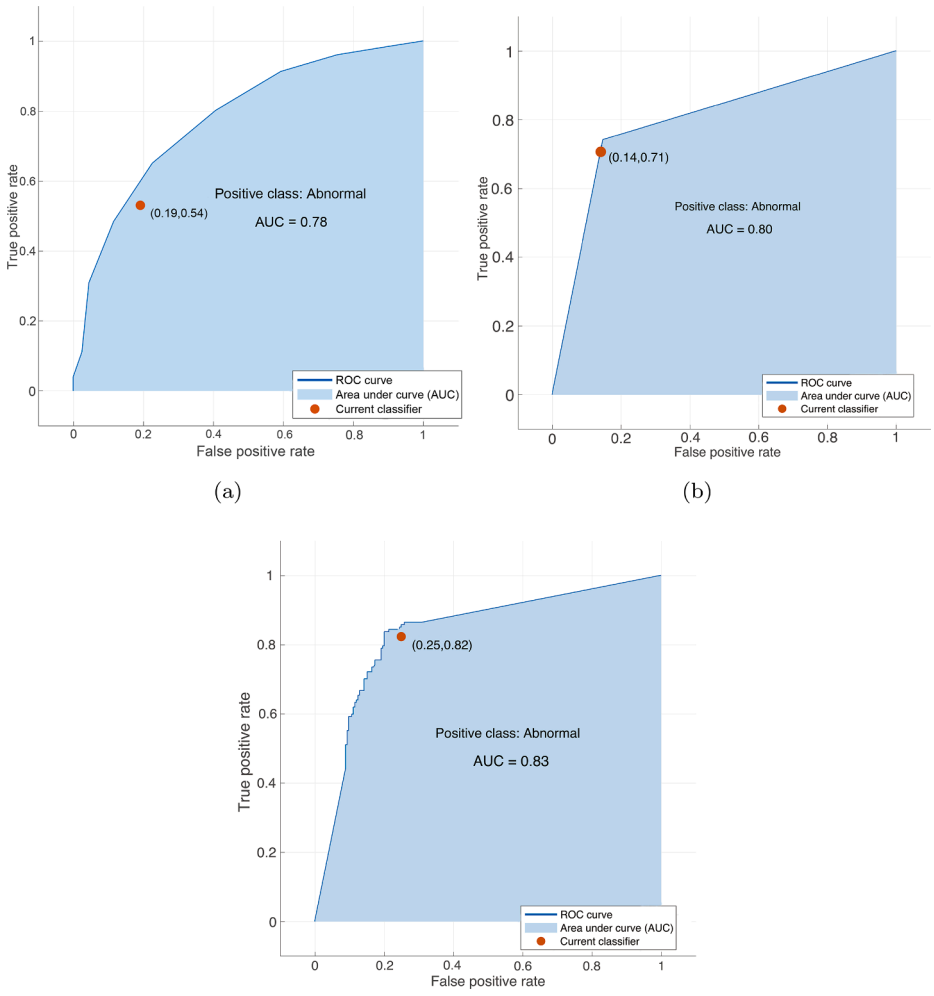


Fig. 11. Validation ROC Curve of (a) Model 1 (b) Model 2 (c) Model 3.

(b) and the calculated AUC value also verified the improvement.

Model 3: By adjusting the weights, the performance is shown in Table 2 under label Model 3, and the Validation ROC is shown in Fig. 11 (c). This classification model provides a substantial improvement in the judgment of positive patients without a significant reduction in accuracy. Although the false-positive component rises slightly, slightly affecting the performance evaluation of the classifier, such a cost should be acceptable in the anticipated application scenario.

3.2.2. *CNN performance*

The performance results are listed in Table 3. From the table, it can be seen that performance by using MFCCs as features have been significantly improved by comparing with taking the raw data directly from the breath sound, where accuracy can reach more than 97%. This verifies the effectiveness of taking MFCCs as features and the appropriate approach of using breath sound to diagnose COVID-19.

In addition, Table 3 shows the model performance when these two pre-processing steps were not-employed, partly-employed or fully-employed. As observed from the table, when using pre-emphasized and normalized 13-dimensional MFCCs as the input features, the proposed CNN model had the best test accuracy, sensitivity, and precision; the AUC is also almost the best, while the specificity is only 0.11% lower than the best. Although this model does not achieve the best AUC and specificity, we still consider it to have the best prediction performance among the four. This is because of two reasons. Firstly, even though the specificity is not the best, it is still higher than 99% and is only 0.11% lower than the best. Secondly, sensitivity should be paid slightly more attention than specificity here because identifying COVID-19 positive patients is more important. Apart from the prediction performance, it can be observed that when using pre-emphasized and normalized 13-dimensional MFCCs as the input features, the model achieves the best-optimized state at the earliest epochs. That is to say, with both pre-

Table 3
Model Performance with Different Pre-processing Steps.

Input Feature	Pre-emphasized	Normalized	Test Accuracy	Optimized Epoch Number	AUC	Sensitivity	Specificity	Precision
Raw Data	N	N	60.85%	84	0.6877	54.45%	67.32%	66.85%
MFCC	N	N	97.51%	90	0.9972	94.10%	99.80%	99.69%
MFCC	Y	N	97.04%	64	0.9969	93.62%	99.59%	99.41%
MFCC	N	Y	97.57%	80	**0.9979**	94.19%	**99.90%**	**99.85%**
MFCC	Y	Y	**97.63%**	**61**	0.9977	**94.55%**	99.80%	99.70%

emphasis and normalization, not only the prediction performance of the model is the best, but also the training process can be the most time-saving.

To further explore the difference between using original, pre-emphasized, and normalized 13-dimensional MFCCs as the input features, training loss and validation loss for both cases were plotted. Comparing Fig. 11(c) and Fig. 12(b), we can find that although the general trend of validation losses is both decreasing, there is a rather significant discrepancy at the early stage of training the CNN model. When using pre-emphasized and normalized MFCCs as the input feature, validation loss can be much higher than the training loss for the first 15 epochs, while this pattern does not show when original MFCCs are fed into the CNN model. This observation is worth highlighting because if the early-stopping strategy is employed for the training process, a distorted conclusion that pre-emphasizing and normalizing the MFCCs are unwanted and unnecessary may be drawn, given that the training would stop well before the neural network is properly optimized for the classification task. Actually, when the pre-processed MFCCs are used as input features, the CNN model needs to be trained for about 20 epochs before a synchronous decrease in training and validation loss can be observed. As shown in Table 3, if the models with original or pre-processed MFCCs as input features are both trained for a sufficiently large number of epochs (100 epochs), it can be observed that pre-processing slightly increases the best-optimized performance of the CNN model.

Table 4 shows the model performance with or without Noise Reduction. It can be found that the noise reduction step does actually improve the overall prediction accuracy. It can be further analyzed from the result that the sensitivity is improved, while the specificity slightly deteriorates. Moreover, with the audio de-noised, the AUC and precision are slightly higher and the proposed CNN model can be optimized at the cost of a similar number of epochs. To conclude, considering that the sensitivity, which is perhaps of the greatest significance for our task, is higher when noise reduction is added. This can be confirmed that noise reduction is necessary for the best-optimized model. However, this result may differ based on the choice of dataset. With the dataset of little noise, this step may be unnecessary and even unwanted, given that noise reduction technique *spectral subtraction* may remove patterns from the feature, which are critical for the classification between unhealthy and healthy users.

Apart from pre-processing with pre-emphasis, normalization, and denoising, we want to explore whether the 3-s audio chunk is the best length for diagnosing COVID-19 or not. Table 5 shows the model performance when audio chunks of different lengths are processed with the MFCC package [34] and then used as input features. As can be observed from this table, when audio chunks of 3 s are fed into the CNN model, the prediction accuracy of the model is the best. Moreover, the sensitivity is the highest, implying that this model has the best ability to distinguish COVID-19 positive patients from the crowd. It can be argued that this time interval reaches the best compromise between the size of the training dataset and the length of MFCC plots. To be specific, the test accuracy is the lowest when 1-s audio chunks are fed into the CNN, which is likely to be caused by the narrow MFCCs that lack the crucial patterns for the CNN to distinguish between healthy and unhealthy samples. On the other hand, although 6-s audio chunks are used to train the model for diagnosing COVID-19 using cough sounds [8], in our experiment, 6-s audio chunks may not be the best choice because the total number of audio chunks for training, validating and testing is less than 3000. As a result of this basic evaluation, it can be concluded that 3-s audio chunks are enough to reveal the different patterns of MFCCs between healthy and unhealthy samples.

There are several other variants of MFCC, including ΔMFCC, ΔΔMFCC, and log-mel spectrum, inspired from MFCC and sometimes they outperform MFCC for sound classification and recognition [37–39]. For this task, we want to explore whether ΔMFCC, ΔΔMFCC, and log-Mel spectrum can be effective features for the classification between healthy and unhealthy samples.

ΔMFCC, and ΔΔMFCC. They are delta-cepstral features proposed in [38] to add dynamic information to the static MFCC features. A previous study in [39] noted that the addition of these features to the static 13-dimensional MFCCs strongly improved speech recognition accuracy. However, there is a lack of research on sound classification tasks.

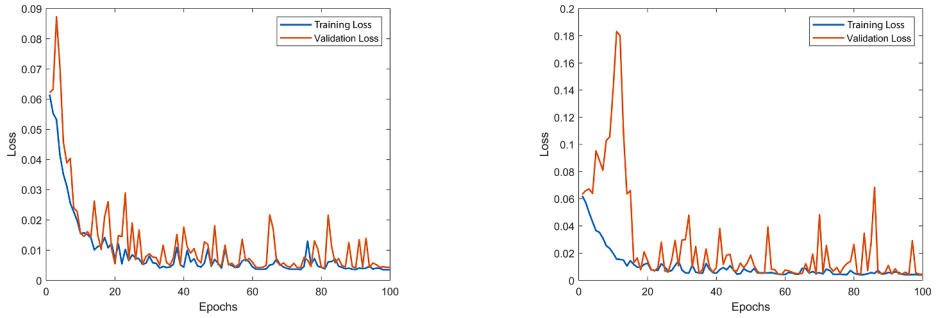
Log-Mel Spectrum. As suggested in [37], log-mel spectrum, which removes the DCT step of deriving MFCCs, is likely to be more appropriate when deep learning is used for digital signal processing because DCT removes information and destroys spatial relations of the breathing signals.

Table 6 displays the model performance when these MFCC-inspired features are used to train the CNN with the length of audio chunks to be 3 s. It can be seen that adding delta-cepstral features to MFCC does not improve the classification accuracy. We further explore whether ΔMFCC or ΔΔMFCC can be used alone as effective features for the classification. As shown in Table 6, the test accuracy of both of them does not remain competitive with that of MFCC. Therefore, it is reasonable to say that adding them to the MFCC impairs the classification effect. On the other hand, the CNN trained with log-Mel spectrum has slightly lower test accuracy than the model trained with MFCCs. This result is contradicting the argument in a previous study [37] that DCT is unwanted and unnecessary with deep learning models processing sound signals.

3.2.3. Android app performance

From the signal analysis and performance evaluation of both classifiers *k*NN and CNN, it can be seen that diagnosis of COVID-19 on smartphones via acoustic analysis is a feasible approach to detect COVID-19. The evaluation of two different classifiers cross-verified the effectiveness of using machine learning methods to support the classification of COVID-19. Between these two classifiers, CNN is apparently a better choice for implementing this diagnosis technique.

Currently, medical trials are being organized. The medical trial will



(a) (b)

Fig. 12. Training Loss and Validation Loss with (a) Original 13-dimensional MFCC, and (b) Pre-emphasized and Normalized 13-dimensional MFCC.

Table 4
Model Performance with or without Noise Reduction.

Pre-emphasized & Normalized	De-noised	Test Accuracy	Optimized Epoch Number	AUC	Sensitivity	Specificity	Precision
Y	N	97.63%	61	0.9977	94.55%	**99.80%**	99.70%
Y	Y	**97.87%**	**59**	**0.9978**	**95.26%**	99.79%	**99.71%**

Table 5
Model Performance with Different Lengths of Audio Chunks (ACs).

Length of ACs (s)	Total Number of ACs	Test Accuracy	AUC	Sensitivity	Specificity	Precision
1	21789	96.04%	0.9944	91.93%	98.99%	98.50%
2	10529	96.70%	0.9964	92.85%	99.36%	99.01%
3	6748	**97.87%**	0.9978	95.26%	99.79%	99.71%
4	4943	95.95%	**0.9986**	**99.80%**	93.39%	90.98%
5	3778	97.57%	0.9982	93.98%	**100.00%**	**100.00%**
6	2970	96.64%	0.9884	98.37%	95.41%	93.79%

Table 6
Model Performance with Different Features (Each Model is Trained 6 Times).

Features	Test Accuracy	AUC	Sensitivity	Specificity	Precision
MFCC	97.87%	0.9978	95.26%	99.79%	99.71%
ΔMFCC	97.04%	0.9980	93.16%	100.00%	100.00%
ΔΔMFCC	97.45%	0.9979	94.45%	99.50%	99.23%
MFCC + ΔMFCC + ΔΔMFCC	97.39%	0.9979	93.61%	100.00%	100.00%
Log-mel Spectrum	97.21%	0.9976	93.75%	99.79%	99.70%

have two purposes; one is to validate the method in diagnosis; the other is to define a probability threshold to confirm the confidence level of the diagnosis. The first test has been completed by following the medical protocol. Test results of this method and lateral flow are shown in Fig. 13. It can be seen that the results are quite similar between these two methods. Details of the medical trial and more results will be reported afterward.

4. Discussion

From the investigation of this diagnosis technique, several solid findings were discovered and lessons were also learned on the possible limitations of this method. These findings include:

(1) Signal analysis of breath sound in both time domain and frequency domain clearly showed that the breath sound after contracting COVID-19 differs scientifically from the healthy people. This has provided a piece of solid scientific evidence for using breath sound to diagnose COVID-19.

(2) Same as other research such as diagnosing COVID-19 from coughing, artificial intelligence plays a very important role in classifying patients with COVID-19 from healthy people.

(3) Initial medical test verified that the proposed diagnosis method in this paper is feasible and effective.

However, during the development of this method, a few issues need to be considered, including:

(1) Dateset: Because of the limited volume of breath sounds data provided in the existing dataset, although some samples of other lung diseases were provided and models were trained to differentiate COVID-19 from these diseases, samples didn't cover the whole range of lung diseases. There is a possibility that some other lung diseases may have similar patterns to COVID-19. Broader research could be conducted to analyze the acoustic properties of all known diseases so as to avoid the wrong diagnosis. Whist this paper is being prepared and revised, a new variant name Omicron was detected in South Africa. A scientific study published in Nature revealed that this variant causes less damage to the lungs than upper airways [40]. If the lung is not infected or infected slightly, in theory, the acoustic signals produced from patients with Omicron might be different from previous variants. In this case, the existing model could be trained again with new datasets to diagnose Covid-19. Research is ongoing with the development of a new variant of Covid-19 and results will be reported once any new results were discovered.

(2) Raw data or pre-processed data for training and testing: From the test results, it can be observed that after signal pre-processing, performance has been significantly improved. This is because signal pre-processing emphasizes the importance of features, which will have a better contribution to the results.

(3) Length of the sound chunks: Although the 3-s sound chunk is supposed to be the best one to produce the highest accuracy, it didn't differ too much from other chunks. The accuracy maintains high for all chunks with different lengths. However, different length of chunks affects the computation speed, particularly on smartphones. This parameter could be well balanced in practical development by balancing the accuracy and computation load.

(4) Feature selection: Selection of the right features is of utmost importance in training the models. Based on the nature of the artificial intelligence algorithms, the right number and type of features may significantly improve the performance. This is the area that should be considered seriously in practical development.

(5) Training of the model: Models cannot be used directly to diagnose COVID-19. They need to be trained and optimized by considering the data type, effective features, amount of data, imbalance of the data,

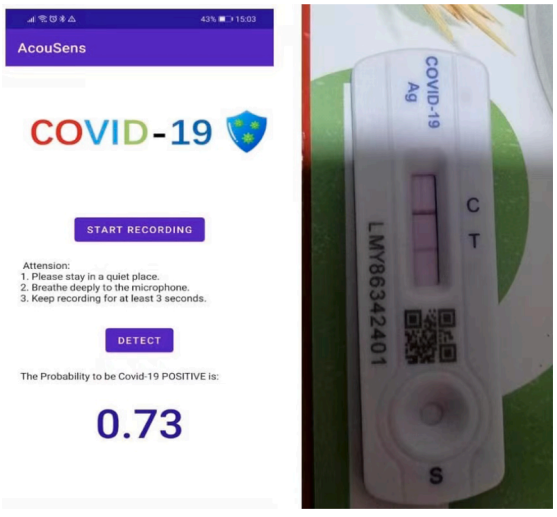


Fig. 13. Test results and lateral flow.

pre-processing of the data, and appropriate training methods to develop a high-performance model for diagnosis.

(6) Complex CNN model can be implemented in smartphones so as to make this diagnosis method applicable to everybody. It is envisioned that this technique will have tremendous impacts on society in fighting against COVID-19. Apart from the aforementioned providing a mass detection method, this technique will particularly help medical professionals to reduce their contact time with patients so as to protect themselves from contracting COVI-19 from the patient, where quite a few cases happened across the world that medical staff died from COVID-19 when they were treating patients with COVID-19. This will reduce their stress and burnout, and improve their mental well-being so that they can provide better services to the public because COVID-19 has severely affected the life of medical staff as demonstrated in the latest studies [41,42]. This technique could also be implemented as a telemedicine scheme for diagnosis, treatment, and post-COVID-19 care, which was suggested in [43].

5. Conclusions

A new method of diagnosing COVID-19 is proposed in this paper, which is based on the acoustic analysis of breath sound and artificial intelligence. From the signal analysis, it discovered the scientific evidence to support this method. The high-performance test results and initial medical tests verify the effectiveness of the proposed method. In addition, this method could be implemented in smartphones, which makes it applicable to everybody with a smartphone, regardless of where he/she is or the existence of medical professionals, because breath sound is natural and no particular training is needed. This method doesn't need medical equipment either. Because of no need for training, medical professionals, and medical equipment, this method could be a mass diagnosis method for society to diagnose and fight against COVID-19. Over time, new virus variants will gradually emerge later, in which case the new dataset will need to be used for new training to ensure the accuracy of the model in diagnosing COVID-19.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors would like to thank Xi'an Jiaotong-Liverpool University for her financial support for this group of students to conduct this research under the project SURF-2021039. The work is also partially supported by Key Program Special Fund in XJTLU under project KSF-E-64, XJTLU Research Development Funding under projects RDF-19–01-14 and RDF-20–01-15, and the National Natural Science Foundation of China (NSFC) under grant 52175030.

References

[1] A. Imran, I. Posokhova, H.N. Qureshi, U. Masood, M.S. Riaz, K. Ali, C.N. John, M. I. Hussain, M. Nabeel, Ai4covid-19: Ai enabled preliminary diagnosis for covid-19 from cough samples via an app, Inform. Med. Unlocked 20 (2020) 100378.
[2] L.S. Wen, Hospitals are overwhelmed because of the coronavirus - here's how to help, https://www.wctrib.com/opinion/5001125-Leana-S.-Wen-Hospitals-are-overwhelmed-because-of-the-coronavirus-%E2%80%94-heres-how-to-help.
[3] S. Najmabadi, J. Root, Coronavirus test results in texas are taking up to 10 days, https://www.kxxv.com/your-hometown/texas/coronavirus-test-results-in-texas-are-taking-up-to-10-days.
[4] M. Asiaee, A. Vahedian-Azimi, S.S. Atashi, A. Keramatfar, M. Nourbakhsh, Voice quality evaluation in patients with covid-19: An acoustic analysis, Journal of Voice.
[5] K.D. Bartl-Pokorny, F.B. Pokorny, A. Batliner, S. Amiriparian, A. Semertzidou, F. Eyben, E. Kramer, F. Schmidt, R. Schönweiler, M. Wehler, et al., The voice of covid-19: Acoustic correlates of infection in sustained vowels, J. Acoust. Soc. Am. 149 (6) (2021) 4377–4383.
[6] B. Stasak, Z. Huang, S. Razavi, D. Joachim, J. Epps, Automatic detection of covid-19 based on short-duration acoustic smartphone speech analysis, J. Healthcare Inform. Res. 5 (2) (2021) 201–217.
[7] M. Faezipour, A. Abuzneid, Smartphone-based self-testing of covid-19 using breathing sounds, Telemedicine and e-Health 26 (10) (2020) 1202–1205.
[8] J. Laguarta, F. Hueto, B. Subirana, Covid-19 artificial intelligence diagnosis using only cough recordings, IEEE Open J. Eng. Med. Biol. 1 (2020) 275–281.
[9] A. Pal, M. Sankarasubbu, Pay attention to the cough: Early diagnosis of covid-19 using interpretable symptoms embeddings with cough sound signal processing, in: Proceedings of the 36th Annual ACM Symposium on Applied Computing, 2021, pp. 620–628.
[10] H. Aygün, A. Apolskis, The quality and reliability of the mechanical stethoscopes and laser doppler vibrometer (ldv) to record tracheal sounds, Appl. Acoust. 161 (2020) 107159.
[11] Y. hui Huang, S. jun Meng, Y. Zhang, S. sheng Wu, Y. Zhang, Y. wei Zhang, Y. xiang Ye, Q. feng Wei, N. gui Zhao, J. ping Jiang, et al., The respiratory sound features of covid-19 patients fill gaps between clinical data and screening methods, medRxiv.
[12] C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, C. Mascolo, Exploring automatic diagnosis of covid-19 from crowdsourced respiratory sound data, arXiv preprint arXiv:2006.05919.
[13] Y. Xu, X. Liu, X. Cao, C. Huang, E. Liu, S. Qian, X. Liu, Y. Wu, F. Dong, C.-W. Qiu, et al., Artificial intelligence: A powerful paradigm for scientific research, The Innovation 2 (4) (2021) 100179.
[14] M. Pahar, M. Klopper, R. Warren, T. Niesler, Covid-19 detection in cough, breath and speech using deep transfer learning and bottleneck features, Comput. Biol. Med. 141 (2022) 105153.
[15] H. Coppock, A. Gaskell, P. Tzirakis, A. Baird, L. Jones, B. Schuller, End-to-end convolutional neural network enables covid-19 detection from breath and cough audio: a pilot study, BMJ innovations 7 (2).
[16] B. Stasak, Z. Huang, S. Razavi, D. Joachim, J. Epps, Automatic detection of covid-19 based on short-duration acoustic smartphone speech analysis, J. Healthcare Inform. Res. 5 (2) (2021) 201–217.
[17] C. Gomes, Report of the who-china joint mission on coronavirus disease 2019 (covid-19), Brazilian Journal of Implantology and Health Sciences 2 (3).
[18] A. Bendix, S. Gal, How omicron symptoms differ from the delta variant and original strain in two charts, https://www.businessinsider.com/omicron-common-symptoms-vs-other-variants-charts-2022-1 (1 2022).
[19] S. Tian, W. Hu, L. Niu, H. Liu, H. Xu, S.-Y. Xiao, Pulmonary pathology of early-phase 2019 novel coronavirus (covid-19) pneumonia in two patients with lung cancer, Journal of thoracic oncology 15 (5) (2020) 700–704.
[20] H.X. Bai, B. Hsieh, Z. Xiong, K. Halsey, J.W. Choi, T.M.L. Tran, I. Pan, L.-B. Shi, D.-C. Wang, J. Mei, et al., Performance of radiologists in differentiating covid-19 from non-covid-19 viral pneumonia at chest ct, Radiology 296 (2) (2020) E46–E54.
[21] N. Sharma, P. Krishnan, R. Kumar, S. Ramoji, S.R. Chetupalli, P.K. Ghosh, S. Ganapathy, et al., Coswara–a database of breathing, cough, and voice sounds for covid-19 diagnosis, arXiv preprint arXiv:2005.10548.
[22] L. Rabiner, B.-H. Juang, Fundamentals of speech recognition, Prentice-Hall Inc, 1993.
[23] A. Hashemi, H. Arabalibiek, K. Agin, Classification of wheeze sounds using wavelets and neural networks, in: International Conference on Biomedical Engineering and Technology, Vol. 11, IACSIT Press, 2011, pp. 127–131.
[24] L. Pesu, P. Helistö, E. Ademovič, J.-C. Pesquet, A. Saarinen, A. Sovijärvi, Classification of respiratory sounds based on wavelet packet decomposition and learning vector quantization, Technol. Health Care 6 (1) (1998) 65–74.
[25] C. Turner, A. Joseph, A wavelet packet and mel-frequency cepstral coefficients-based feature extraction method for speaker identification, Procedia Computer Science 61 (2015) 416–421.
[26] S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. Acoust., Speech, Signal Process. 28 (4) (1980) 357–366.
[27] L.-S.A. Low, N.C. Maddage, M. Lech, L.B. Sheeber, N.B. Allen, Detection of clinical depression in adolescents' speech during family interactions, IEEE Trans. Biomed. Eng. 58 (3) (2010) 574–586.
[28] X. Wang, J. Zhang, Y. Yan, Discrimination between pathological and normal voices using gmm-svm approach, J. Voice 25 (1) (2011) 38–43.
[29] M.A.R. Díaz, C.A.R. García, L.C.A. Robles, J.E.X. Altamirano, A.V. Mendoza, Automatic infant cry analysis for the identification of qualitative features to help opportune diagnosis, Biomed. Signal Process. Control 7 (1) (2012) 43–49.
[30] H. Mansy, T. Royston, R. Balk, R. Sandler, Pneumothorax detection using computerised analysis of breath sounds, Med. Biol. Eng. Comput. 40 (5) (2002) 526–532.
[31] M.K. Arjmandi, M. Pooyan, M. Mikaili, M. Vali, A. Moqarehzadeh, Identification of voice disorders using long-time features and support vector machine with different feature reduction methods, J. Voice 25 (6) (2011) e275–e289.
[32] S. Matos, S.S. Birring, I.D. Pavord, H. Evans, Detection of cough signals in continuous audio recordings using hidden markov models, IEEE Trans. Biomed. Eng. 53 (6) (2006) 1078–1083.
[33] M.V. Valueva, N. Nagornov, P.A. Lyakhov, G.V. Valuev, N.I. Chervyakov, Application of the residue number system to reduce hardware costs of the convolutional neural network implementation, Math. Comput. Simul. 177 (2020) 232–243.
[34] B. McFee, M.M.S. Balke, C. Thomé, C. Raffel, D. Lee, O. Nieto, E. Battenberg, D. Ellis, R. Yamamoto, J. Moore, R. Bittner, K. Choi, P. Friesch, F.-R. Stöter, V. Lostanlen, S. Kumar, S. Waloschek, Seth, R. Naktinis, D. Repetto, C.F. Hawthorne, C. Carr, W. Pimenta, P. Viktorin, P. Brossier, J.F. Santos, JackieWu, Erik, A. Holovaty, librosa/librosa: 0.6.1, 2018, doi: 10.5281/zenodo.1252297.

[35] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, P.J. Kennedy, Training deep neural networks on imbalanced data sets, in: 2016 international joint conference on neural networks (IJCNN), IEEE, 2016, pp. 4368–4374. doi:10.1109/IJCNN.2016.7727770.
[36] X. Wang, Diagnosis of covid-19 on smartphone within 2 minutes, https://www.acousens.care/.
[37] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, T. Sainath, Deep learning for audio signal processing, IEEE J. Select. Top. Signal Process. 13 (2) (2019) 206–219.
[38] S. Furui, Speaker-independent isolated word recognition based on emphasized spectral dynamics, in: ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 11, IEEE, 1986, pp. 1991–1994.
[39] K. Kumar, C. Kim, R.M. Stern, Delta-spectral cepstral coefficients for robust speech recognition, in: 2011 IEEE International conference on acoustics, speech and signal processing (ICASSP), IEEE, 2011, pp. 4784–4787.
[40] K. M, Omicron's feeble attack on the lungs could make it less dangerous, https://www.nature.com/articles/d41586-022-00007-8, 2022, doi:10.1038/d41586-022-00007-8 (1.
[41] T.G. Kannampallil, C.W. Goss, B.A. Evanoff, J.R. Strickland, R.P. McAlister, J. Duncan, Exposure to covid-19 patients increases physician trainee stress and burnout, PloS one 15 (8) (2020) e0237301.
[42] B.A. Evanoff, J.R. Strickland, A.M. Dale, L. Hayibor, E. Page, J.G. Duncan, T. Kannampallil, D.L. Gray, Work-related and personal factors associated with mental well-being during the covid-19 response: survey of health care and other workers, J. Med. Internet Res. 22 (8) (2020) e21366.
[43] T. Kannampallil, J. Ma, Digital translucence: adapting telemedicine delivery post-covid-19, Telemedicine and e-Health 26 (9) (2020) 1120–1122.