Available at www.sciencedirect.com

ScienceDirect

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

Original Research Article

Explainable COVID-19 detection using fractal dimension and vision transformer with Grad-CAM on cough sounds

Nebras Sobahi[a,*], *Orhan Atila*[b], *Erkan Deniz*[b], *Abdulkadir Sengur*[b], *U. Rajendra Acharya*[c,d,e]

[a] *King Abdulaziz University, Department of Electrical and Computer Engineering, Jeddah, Saudi Arabia*
[b] *Firat University, Technology Faculty, Electrical and Electronics Engineering Department, Elazig, Turkey*
[c] *Ngee Ann Polytechnic, Department of Electronics and Computer Engineering, Singapore*
[d] *Biomedical Engineering, School of Science and Technology, SUSS University, Singapore*
[e] *Biomedical Informatics and Medical Engineering, Asia University, Taichung, Taiwan*

ARTICLE INFO

Article history:
Received 23 May 2022
Received in revised form
28 August 2022
Accepted 29 August 2022
Available online 6 September 2022

Keywords:
Cough sound
COVID-19 detection
YAMNet
Fractal dimension
Vision Transformer

ABSTRACT

The polymerase chain reaction (PCR) test is not only time-intensive but also a contact method that puts healthcare personnel at risk. Thus, contactless and fast detection tests are more valuable. Cough sound is an important indicator of COVID-19, and in this paper, a novel explainable scheme is developed for cough sound-based COVID-19 detection. In the presented work, the cough sound is initially segmented into overlapping parts, and each segment is labeled as the input audio, which may contain other sounds. The deep Yet Another Mobile Network (YAMNet) model is considered in this work. After labeling, the segments labeled as cough are cropped and concatenated to reconstruct the pure cough sounds. Then, four fractal dimensions (FD) calculation methods are employed to acquire the FD coefficients on the cough sound with an overlapped sliding window that forms a matrix. The constructed matrixes are then used to form the fractal dimension images. Finally, a pretrained vision transformer (ViT) model is used to classify the constructed images into COVID-19, healthy and symptomatic classes. In this work, we demonstrate the performance of the ViT on cough sound-based COVID-19, and a visual explainability of the inner workings of the ViT model is shown. Three publically available cough sound datasets, namely COUGHVID, VIRUFY, and COSWARA, are used in this study. We have obtained 98.45%, 98.15%, and 97.59% accuracy for COUGHVID, VIRUFY, and COSWARA datasets, respectively. Our developed model obtained the highest performance compared to the state-of-the-art methods and is ready to be tested in real-world applications.

* Corresponding author at: King Abdulaziz University, Department of Electrical and Computer Engineering, Jeddah, Saudi Arabia
E-mail addresses: nsobahi@kau.edu.sa (N. Sobahi), oatila@firat.edu.tr (O. Atila), edeniz@firat.edu.tr (E. Deniz), ksengur@firat.edu.tr 1 (A. Sengur), Rajendra_Udyavara_ACHARYA@np.edu.sg (U.R. Acharya).

https://doi.org/10.1016/j.bbe.2022.08.005
0168-8227/© 2022 Nalecz Institute of Biocybernetics and Biomedical Engineering of the Polish Academy of Sciences. Published by Elsevier B.V. All rights reserved.

1. Introduction

As the COVID-19 virus has become a pandemic, it is important to determine a proper way to determine if people are infected or not [1,2]. To this end, the polymerase chain reaction (PCR) test has become a popular method to determine the virus [3]. Although the PCR test is the most used COVID-19 detection method, the performance of the test is not yet at the desired level. In this context, it is critical to examine the use of commonplace instruments such as smartphones and machine learning as methods for detecting COVID-19 infection as a way of reducing the need for every-one to do tests. Coughing is a typical symptom of several upper respiratory disorders, including asthma, bronchitis, pertussis, and COVID-19 [4]. The coughing sound is usually distinct for each respiratory ailment, allowing doctors to diagnose the sickness based on the cough sound alone.

Although most of the artificial intelligence-based COVID-19 detection studies were either based on the analysis of chest X-ray or CT images, cough sound analysis-based COVID-19 detection has also gained much attention in the artificial intelligence and signal processing communities [4]. Pahar et al. [5] presented an artificial intelligence-based approach for detecting COVID-19 based on the cough audios recorded by smartphones. The authors used two datasets in their study, and the minority oversampling technique was considered to alleviate the dataset skew problem. Seven machine learning approaches were used in the classification stage of the study, and the cross-validation technique was considered in the performance evaluation. The authors reported that the best achievement was obtained by using the ResNet50 classifier, where the calculated AUC was 0.98. Laguarta et al. [6] proposed a machine learning-based speech processing approach for cough recordings-based COVID-19 detection. The input cough audios were initially converted to the cough images by using the Mel Frequency Cepstral Coefficient (MFCC). The obtained images were then fed into the Convolutional Neural Networks (CNN) based architecture. Authors reported 98.5 % sensitivity and 94.2 % specificity scores. Alsabek et al. [7] presented a study that employed cough and breathing sounds and speech audios to diagnose COVID-19. For feature extraction, the authors used the MFCCs, and for classification, they used Pearson's Correlation coefficient values. Sharma et al. [8] compiled the Coswara dataset, which comprises audio recordings of cough, breath, and speech. The authors used the random forest method as a classifier and extracted 28 spectral characteristics. The overall accuracy score was reported to be 67.7 %. Mouawad et al. [9] suggested an approach involving cough and other vocal audios. The authors retrieved the MFCC features in the classification phase and employed various machine learning approaches: decision trees, k-nearest neighbor, random forest, and XGBoost. The authors reported a 97.0 percent accuracy rate and a 62.0 percent F1-score with the XGBoost classifier. Erdogan et al. [10] proposed an approach where cough sounds were used to detect COVID-19. In this context, the authors used the multiresolution approaches for feature extraction, namely empirical-mode and wavelet decompositions. Moreover, various pre-trained CNN models were also used for feature extraction. All these features were then concatenated, and a feature selection mechanism was employed. Lastly, SVM was used for classification purposes. The authors reported a 97.8 % accuracy and a 98.0 % F1-score. Despotovic et al. [11] demonstrated early COVID-19 identification findings utilizing typical acoustic feature sets, wavelet scattering features, and deep audio embedding taken from low-level feature representations. The created models had an accuracy of 88.52 %, a sensitivity of 88.75 %, and a specificity of 90.87 %, demonstrating that audio characteristics may be used to detect COVID-19 symptoms. Tena et al. [12] employed a supervised machine-learning algorithm to propose an automated feature extraction technique based on time–frequency cough characteristics and selecting the more relevant ones to diagnose COVID-19. The authors employed a random forest classifier with nearly 90 % accuracy. Kobat et al. [13] used a graph-based local feature generator, an iterative maximum relevance minimal redundancy iterative feature selector, and the k-nearest neighbor classifier for cough sound-based COVID-19, heart failure, and healthy subject discrimination. For the COVID-19 vs healthy, heart failure vs healthy, and COVID-19 vs heart failure vs healthy classes, the authors reported accuracies of 100.0 %, 99.38 %, and 99.49 %, respectively. Chang et al. [14] developed a viable model that used the large-scale multi-sound FluSense dataset to assist in COVID-19 identification from cough noises. The transfer learning methodology was created and utilized rather than merely augmenting the training data with FluSense due to the gap between FluSense and the COVID-19-related datasets comprising cough exclusively. To this goal, four pretrained CNN models were utilized, and the suggested technique improved the area under the receiver operating characteristic curve by 3.57 percent over the baseline of DiCOVA Track-1 validation. Chowdhury et al. [15] suggested an ensemble-based multi-criteria decision-making strategy for selecting the best COVID-19 cough classification machine learning methodology. To test their strategy, the authors employed four cough datasets: Cambridge, Coswara, Virufy, and NoCoCoDa. After extracting audio attributes from them, the authors used machine learning approaches to categorize the cough samples as COVID-19 or non-COVID-19. The optimal model was then chosen using ensemble technologies using a multi-criteria decision-making process. They have reported an area under curve (AUC) of 95 %, precision of 100 %, and 97 % recall score of 97 % using the extra-trees classifier. Islam et al. [16] used both time and frequency domain features to discriminate the healthy and COVID-19 cough signals. Authors used zero crossing rate, energy, and entropy features. The deep NN was considered as a classification tool. The experiments on Virufy dataset yielded a 93.8 % accuracy score for mixed features from both the time and frequency domain. Hamdi et al. [17] used an attention mechanism-based CNN-LSTM approach for cough-based COVID-19 detection. Authors used a spectral-based data augmentation approach which was based on pitch shifting. The COUGHVID dataset was used in experiments, and the authors reported 91.13 % classification accuracy. Manshouri [18] used spectral analysis and SVM classifier for diagnosis of COVID-19 from cough signals. The author used power spectral density, STFT, and MFCC as features and radial basis function-based SVM classifier on the VIRUFY dataset, and a 95.86 % accuracy score was reported. Lee et al. [19] used mel-scaled spectrogram images and MFCC images with a CNN approach for efficient COVID-19 detection from the cough sound signals. Authors used various pretrained deep CNN models in their works and obtained a 97.2 % accuracy score. Dang et al. [20] developed deep learning-based COVID-19 detection based on cough, breathing, and voice. Gated recurrent units classifier was used in the proposed method. Authors yielded an AUROC of 0.79 value in their experimental works.

Rahman et al. [21] developed a deep CNN approach for cough sound-based COVID-19 detection. Authors used a stacking procedure where eight pretrained CNN models were used. The proposed stacking procedure was based on a logistic regression (LR) classifier. The spectrogram images of the input cough sounds were used as input to the CNN stacker model. The combinations of the Cambridge and the Qatari datasets were used, and a 96.5 % accuracy score was obtained using 5 fold cross validation test. Ren et al. [22] used a set of 6,373 acoustic features based on COMPARE analysis to detect COVID-19 based on cough sound analysis. Authors used various machine learning methods to classify the acoustic features and explored that MFCC and bear-essential-acoustic information were quite efficient in COVID-19 detection. Authors used the COUGHVID dataset in their works and obtained a 0.632 average recall score with 5 fold cross validation test. A comprehensive study was carried out by Sharan et al. [23] to look at the different approaches to cough-based COVID-19 detection. The author investigated if artificial intelligence might be used to differentiate between different cough types. Using the Google Scholar, PubMed, and MIT library search engines, a thorough search was conducted to locate literature relevant to cough detection, discrimination, and epidemiology. Gabaldón-Figueira et al. [24] used an autoregressive moving average analysis for cough-based COVID-19 detection. Authors used the correlation between the changes in cough frequency and COVID-19 incidence. The strength of the correlation was determined by calculating its autocorrelation function. A linear regression approach was used for prediction purposes. Authors used their dataset in their experimental works and obtained satisfactory results. Andreu-Perez et al. [25] proposed a cough analysis system that was based on a clinically validated dataset. A generic scheme that was based on deep learning was developed where EMD was considered for cough sound detection, and deep CNN handled the classification issue by using a tensor of audio sonography. Authors used their datasets and obtained a sensitivity of 96.43 % ± 1.85 % and a specificity of 96.20 % ± 1.74 % scores, respectively. Zealouk et al. [26] used a Hidden Markov Model-based approach for cough sound-based COVID-19 detection. The authors used 5 Hidden Markov Model (HMM) states and eight Gaussian mixture distributions in the proposed model. Besides, MFCC features were also considered to obtain the feature vector. The authors used their datasets and obtained 93.33 % and 86.66 % accuracy scores for healthy and patients, respectively.

According to the reviewed literature, cough sound has a high potential for touchless COVID-19 detection, which might save the person who gets the sample for PCR testing. Furthermore, the studied literature revealed that the deep CNN techniques were the general tendency. The main objective of this work is to develop an efficient approach for cough sound-based COVID-19 detection. To this end, a compact system depicted in Fig. 1 was proposed. The main contributions of the proposed system were:

1. The FD approach was used for converting the cough sounds to cough images.
2. Best of our knowledge, the ViT approach was firstly considered in cough sound-based COVID-19 detection.

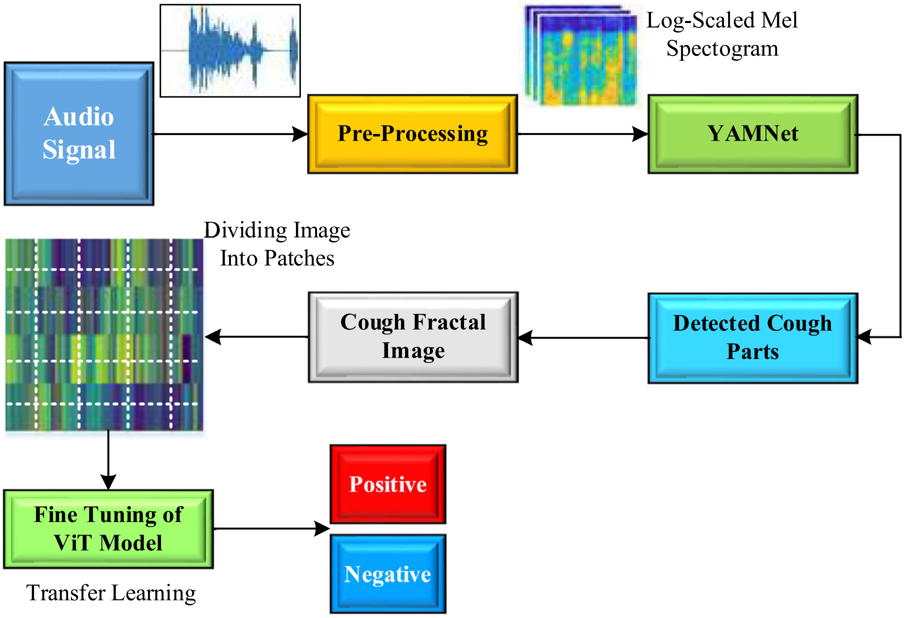


Fig. 1 – Illustration of the proposed method.

3. To the best of our knowledge, the results obtained were higher than those obtained in the literature.

The input sound signals are initially preprocessed for noise removal and amplitude normalization [12]. Windowed median filtering is used in preprocessing stage of the proposed study. As the input cough sound signals contain silence, speech, cough, etc., parts, the YAMNet model is used to detect the cough parts of the input sound signals [27]. And the detected cough parts are used to form a new sound signal where the other sound modalities are alleviated. The cough images fed into the ViT model are obtained using a sliding windowed fractal dimension extraction method. Four fractal dimensions were used to achieve this. The fractal dimensions are calculated using windowed Higuchi, Katz, Castigloni, and Petrosian fractal dimension techniques [28]. The cough images are formed by concatenating the calculated FD coefficients in a matrix. The obtained cough fractal images further train a pretrained ViT model in a transfer learning fashion. Three cough sound datasets, namely COUGHVID, COSWARA, and VIRUFY, are used in experimental works, and various evaluation metrics are employed for performance evaluation. The obtained results show that the proposed method has a huge impact on COVID-19 detection.

2. Related theories

2.1. *Deep audio detector (YAMNet)*

The cough samples detected in the raw audio recordings are automatically identified using the YAMNet [27]. YAMNet is a deep CNN model where the MobileNet [29] was used to classify audio segments into sound classes defined by the AudioSet ontology [30]. YAMNet is composed of 86 layers. These layers are input, 14 convolutional, 13 depth-wise convolutional, 27 ReLu, 27 batch normalization, 1 global pooling, 1 fully connected layer, 1 softmax, and 1 classification output layer, respectively. Fig. 2 shows the implementation of the YAMNet.

As seen in Fig. 2, the input audio signal is pre-processed initially. The input audio signals were buffered into L overlapping segments and resampled to 16 kHz in the preprocessing. Each segment lasted 0.98 s, and the segments were 0.8575 s apart. They were transformed into a magnitude spectrogram with 257 frequency bins using a one-sided short-time Fourier transform using a 25-ms periodic Hann window with a 10-ms hop and a 512-point Discrete Fourier Transform. After that, the magnitude spectrum was passed through a 64-band Mel-spaced filter bank, with the magnitudes of each band combined together. The audio was represented by a 96-by-64-by-1-by-L array, with 96 being the number of spectrums in the Mel spectrogram and 64 being the number of mel bands. Finally, the mel spectrograms were given a log scale. A 96-by-64-by-1-by-L array of mel spectrograms served as the input layer for YAMNet. The output of YAMNet (an L-by-512 matrix) corresponds to confidence scores for each of the 521 sound classifications over time.

2.2. *Vision transformer (ViT)*

The illustration of the ViT structure is given in Fig. 3. The ViT is an image classification model that uses a transformer-like design to classify image patches [31]. An image is divided into fixed-size patches, which are then linearly embedded, position embeddings added, and the resultant vector sequence is given to a conventional transformer encoder. Finally, the traditional strategy of adding an extra learnable “classification token” to the sequence is employed to do classification.

2.3. *Fractal dimension (FD)*

The nonlinear approach of FD based feature extraction is employed to describe the non-regular and self-similarities in a given signal. For many years, FD has been utilized to analyze biological signals like the electroencephalogram and magnetoencephalogram. FD is now being actively used in fields such as speaker identification, sound activity recognition, and especially speech-based emotion classification [28].

2.3.1. *Katz FD*

The Katz FD for a given signal that has the number of n samples is calculated as follows [32]:

$$L = \sum_{i=1}^{n-1} \sqrt{(y_{i+1} - y_i)^2 + (x_{i+1} - x_i)^2} \tag{1}$$

$$d = \max(\sqrt{(y_i - y_1)^2 + (x_i - x_1)^2} \tag{2}$$

where *d* indicates the max distance between the point (x_1, y_1) and the other points, and *L* shows the sum of the distances between neighbor points. Lastly, the Katz FD can be written as:

$$KatzFD = log10(n-1)/(log10(n-1) + log10(d/L)) \tag{3}$$

2.3.2. *Higuchi FD*

Higuchi FD is an iterative approach for analyzing a digital signal for various time scales. This implies that the signal is sampled at various frequencies for a given continuous signal.

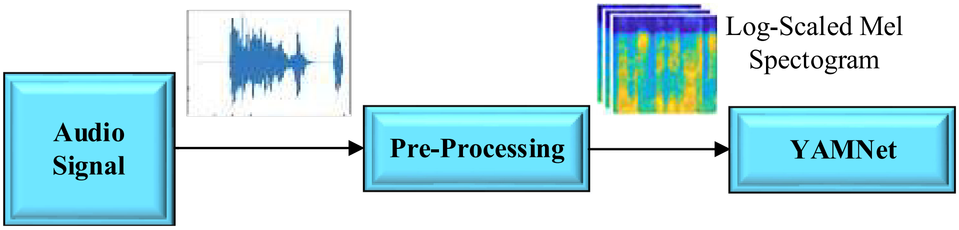


Fig. 2 – Implementation of the YAMNet.

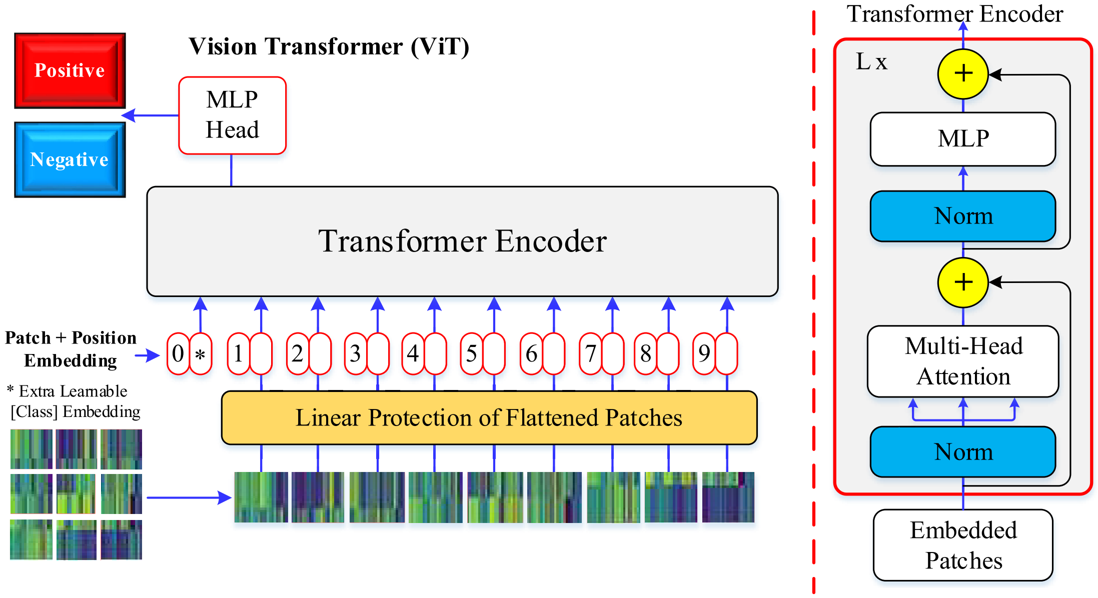


Fig. 3 – Illustration of the ViT [31].

Using several time scales can reveal the signal's frequency features [33]. The new time series is stated as follows if we reconstruct the time series:

$$x_n^l = \left\{x(n), x(n+l), x(n+2l), \ldots, x\left(n + \frac{m-n}{l} l\right)\right\},$$
$$n = 1,\ 2\ \ldots,\ l, \quad l = 1, 2 \ldots,\ l_{\max} \tag{4}$$

where l_{max} is set to 5 for this investigation, n indicates the time index of the signal, l is the resampling range, and s is the integer component of the ratio s. $L_m(l)$, the normalized cumulative change of the signal, was then computed, and $L(l)$, the mean value of $L_m(l)$, is shown as follows:

$$L(l) = \frac{1}{l}\sum_{n=1}^{l} L_n(l) \tag{6}$$

Thus, the Higuchi FD can be determined as the slope of $L(l)$ versus $1/l$ as follows:

$$HFD = \frac{\ln(L(l))}{\ln\left(\frac{1}{l}\right)} \tag{7}$$

2.3.3. *Petrosian FD*

Petrosian FD is one of the simple FD approaches where its computational complexity is less than the others [34]. In Petrosian FD calculation, let's consider a time series $x_1,\ x_2, \ldots, x_N$ and its corresponding waveform points$\{y_1,\ y_2,\ \ldots,\ y_N\}$. Thus, the binary matrix z_i is calculated as follows:

$$z_i = \begin{cases} 1 & x_i > mean(y) \\ -1 & xi \leqslant mean(y) \end{cases},\ i = 1, 2, 3 \ldots, n \tag{8}$$

After calculation of the binary matrix z_i, the changed number of adjacent values in the time series is obtained via Equation (9):

$$n_\Delta = \sum_{i=1}^{n-2}\left|\frac{z_{i+1} - z_i}{2}\right| \tag{9}$$

Finally, the Petrosian FD is calculated as:

$$PFD = \frac{\log 10(n)}{\log 10\left(\frac{n}{n+0.4n_\Delta}\right)} \tag{10}$$

2.3.4. *Castigloni FD*

Castigloni FD is a modified version of Katz FD [35]. The L and d values are calculated as follows:

$$L = \sum_{i=1}^{n-1}\sqrt{(y_{i+1} - y_i)^2 + (x_{i+1} - x_i)^2} \tag{11}$$

$$d = \max(y_i) - \min(y_i) \tag{12}$$

where L shows the total distances between the neighbor points, and d shows the non-stable range of the input time series. As seen in Equations (11) and (12), the calculation of the L value is identical in Katz FD, but the calculation of the d value is different. Thus, the Castigloni FD is calculated as follows:

$$CFD = \frac{\log 10(n-1)}{\log 10(n-1) + \log 10\left(\frac{d}{L}\right)} \qquad .(13)$$

3. Data used

3.1. *COUGHVID dataset*

COUGHVID [36] used a web application to gather cough sounds from 1 April 2020 to 10 September 2020. The user's age, gender, and present condition were input after recording the cough sound. COVID-19, symptomatic, or healthy was the state of the cough sound data. COUGHVID data is in the. webm or.ogg format, with a sampling rate of 48 kHz. The recording lasted at least 2 to 9 s.

3.2. VIRUFY dataset

Stanford University developed the VIRUFY dataset by using a smartphone application [37]. A total of 1187 volunteers were represented in the data. The findings of the RT-PCR test were used to identify which data was positive and which was negative. In this dataset, 595 COVID-19 and 592 healthy volunteers were used.

3.3. COSWARA dataset

Cough recordings were obtained from the general population in the COSWARA dataset using a web-based data collection system and smartphones [38]. The auditory data collected included fast and slow breathing, deep and shallow coughing, phonation of extended vowels, and spoken numbers. Age, gender, location, current health status, and pre-existing medical conditions are all considered. Health status can be described as 'healthy,' 'exposed,' 'cured,' or 'infected.' Audio recordings were recorded at 44.1 kHz.

4. Experimental works

In experiments, both Python and Matlab software were used. The signal processing parts of the proposed method were implemented in Matlab and FD image construction, and ViT model training was implemented in Python. As the input signals contained not only cough segments, the YAMNET model was employed to detect the cough segments. A segmented and classified sound signal is shown in Fig. 4, which contains silence, cough, and speech segments.

As we were interested in cough segments, the detected cough segments were cropped and concatenated to construct the final cough signals. Fig. 5 shows the mentioned process for an input sound signal.

As seen in Fig. 5 (a), the YAMNET was employed to detect the sound events in the given sound signal, and the segments labeled as cough were cropped and concatenated for obtaining the final cough sound signal given in Fig. 5 (b). To be compatible with the YAMNET model, all input audio signals were resampled to 16 kHz. The related other parameters with YAMNET were already given in Section 2.1.

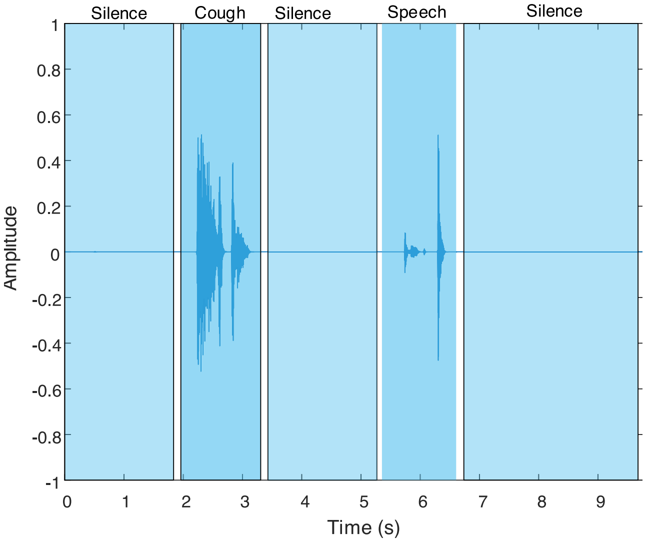


Fig. 4 – A classified sound signal by YAMNET model.

A window of 512 samples with 400 overlapping samples was utilized to calculate FD. The Hanning window is considered in the sliding overlapping window. Fig. 6 shows the constructed FD images. After fractal coefficients are calculated, a normalization process is employed on the fractal coefficients. After concatenation of normalized (0–1 interval) fractal coefficients from all fractal methods, the input image is obtained. It can be noted from Fig. 6 that the Katz, Castigloni, Higuchi, and Petrosian FD were calculated and concatenated to obtain the FD image. The obtained images are then resized to 224 × 224.

As it was shown in Fig. 6, in the FD images, while the x-axis shows the number of segmented signal parts, the y-axis indicates the fractal methods. Fig. 7 shows the constructed FD images for the COUGVID dataset. While the first row of Fig. 7 shows constructed FD images for the COVID-19 class, the second and the third rows show the healthy and symptomatic classes, respectively. For the ViT architecture, we used the ViT-base-patch16-384 model implemented in Python [39]. The patch and the batch sizes were set to 16, and the learning rate was set to 0.0002. The max epoch size was set to 10. During the training of the ViT model, we used the transfer learning strategy where a pretrained ViT model was considered and further trained for detecting the COVID-19 cough sounds [40,41]. In all experimental works, a 10-fold and 5-fold cross-validation evaluation technique were used, and average accuracy, sensitivity, specificity, and F1-score metrics were used for quantitative performance measurements. For COSWARA and VIRUFY datasets, there were two classes (COVID-19 and healthy), and for the COUGHVID dataset, there were three (COVID-19, healthy, and symptomatic) classes.

Table 1 shows the obtained results for the COUGHVID dataset.

As seen in Table 1, we have obtained 98.45 % accuracy, 97.40 % sensitivity, 98.66 % specificity, and 97.70 % F1-score values using the COUGHVID dataset. The cumulative confusion matrix for the COUGHVID dataset is shown in Fig. 8.

As seen in Fig. 8, 614, 6500, and 6142 samples from COVID-19, healthy and symptomatic classes, respectively were correctly classified. Besides, 20 samples from the healthy class were predicted as COVID-19. For the healthy class, 25 samples from COVID-19 and 49 samples from the symptomatic classes were identified as healthy samples. Lastly, only 44 samples from the healthy class were classified as a symptomatic class. Graph of accuracy (%) obtained scores for each fold using the COUGHVID dataset is shown in Fig. 9. As given in Fig. 9, the accuracy values are varied in the range of 97.5 % and 99.5 %.

We have also performed our study using the VIRUFY dataset, and obtained evaluation metrics are given in Table 2. It can be noted from Table 2 that 98.15 % accuracy, 97.48 % sensitivity, 98.82 % specificity, and 98.14 % F1-score values are obtained for the VIRUFY dataset. In addition, the cumulative confusion matrix obtained for the VIRUFY dataset is given in Fig. 10.

As seen in Fig. 10, 580 and 585 samples from COVID-19 and healthy classes, respectively are correctly classified by the

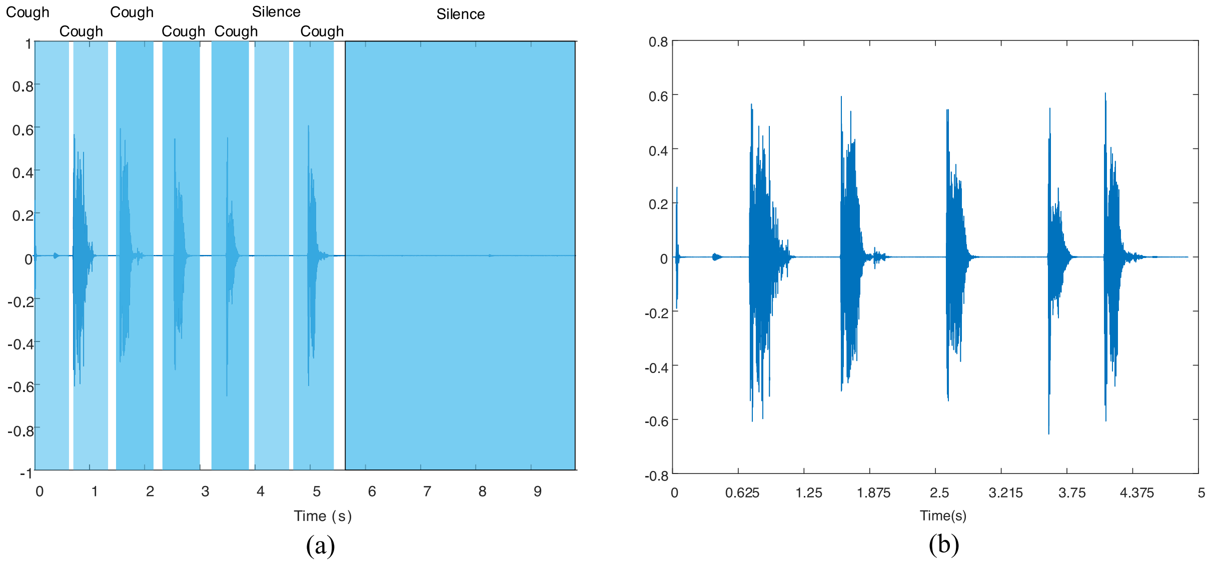


Fig. 5 – Cough sound segmentation process: (a) Segmentation and detection of the sound events; (b) Concatenation of the detected cough segments.

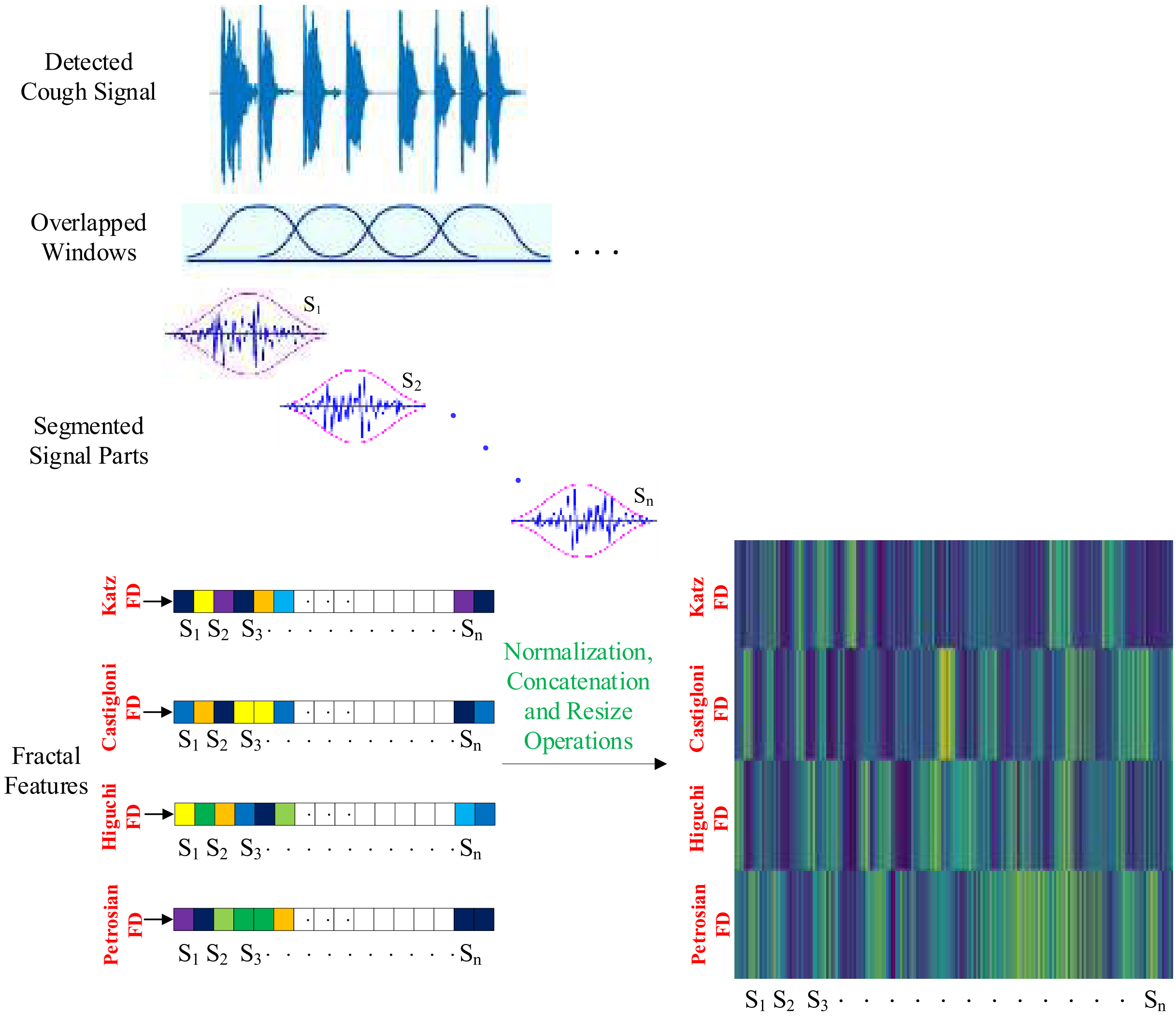


Fig. 6 – Constructed FD sample images.

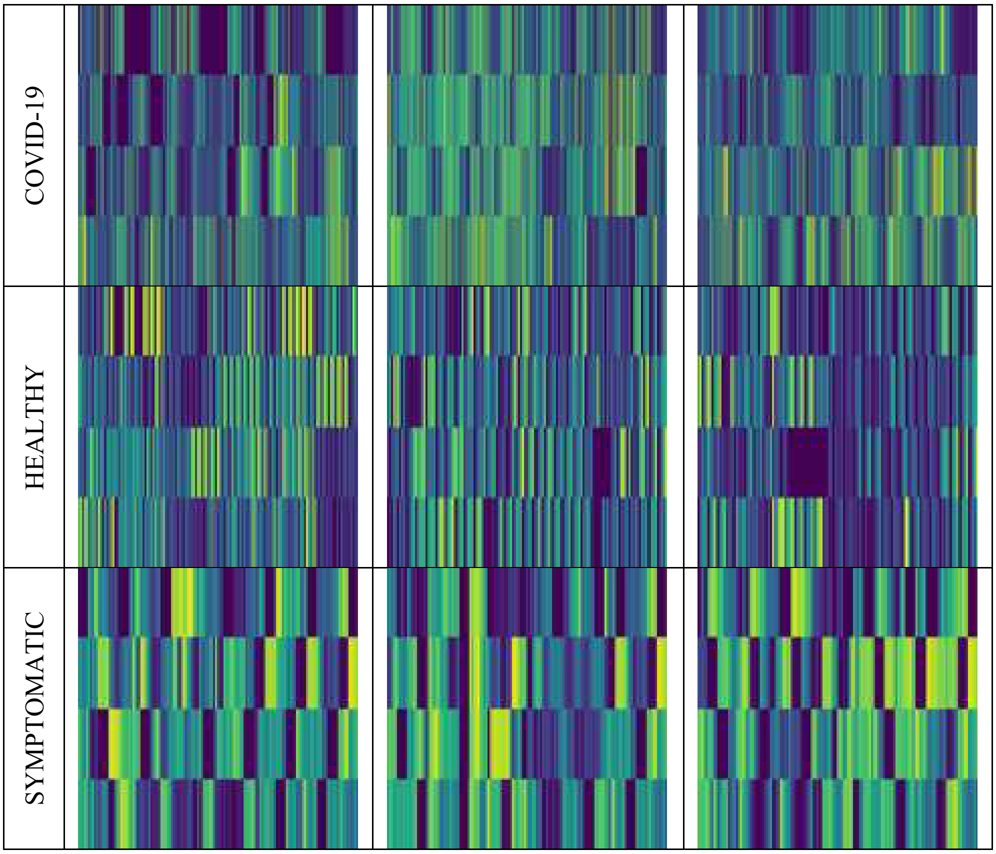


Fig. 7 – Constructed FD sample images for the COUGHVID dataset.

Table 1 – Performance evaluation scores for the COUGHVID dataset.

	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)
COUGHVID	98.45	97.40	98.66	97.70

proposed method. And the proposed scheme missed 15 and 7 samples from COVID-19 and healthy classes, respectively. The graph of accuracy (%) obtained scores for each fold using the VIRUFY dataset is shown in Fig. 11. As seen in Fig. 11, the accuracy scores varied between 97 % and 99 %.

We have also performed our study using the COSWARA dataset, and the outputs obtained are given in Table 3, Fig. 12, and Fig. 13.

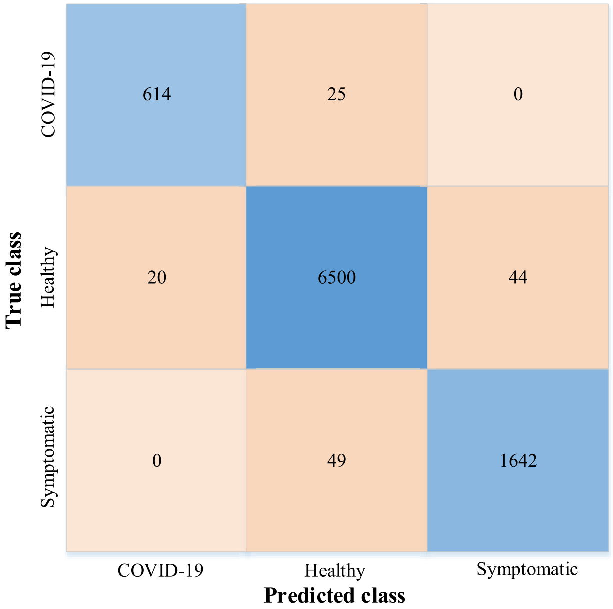
As seen in Table 3, the obtained accuracy, sensitivity, specificity, and F1-score values were 97.59 %, 88.44 %, 98.73 %, and 89.04 %, respectively. When these metrics were compared with the previous evaluation metrics (COUGHVID and VIRUFY dataset), it obtained lower sensitivity and F1-score values.

The cumulative confusion matrix for the COSWARA dataset is also shown in Fig. 12. Again, 130 and 1167 samples are correctly classified in COVID-19 and healthy classes, respectively. And only 17 and 15 samples from COVID-19 and healthy classes were wrongly classified, respectively.

Finally, the accuracy scores obtained for each fold is shown in Fig. 13 for the COSWARA dataset. From this figure, it was obvious that the accuracy scores range between 94.5 % and 99.5 %.

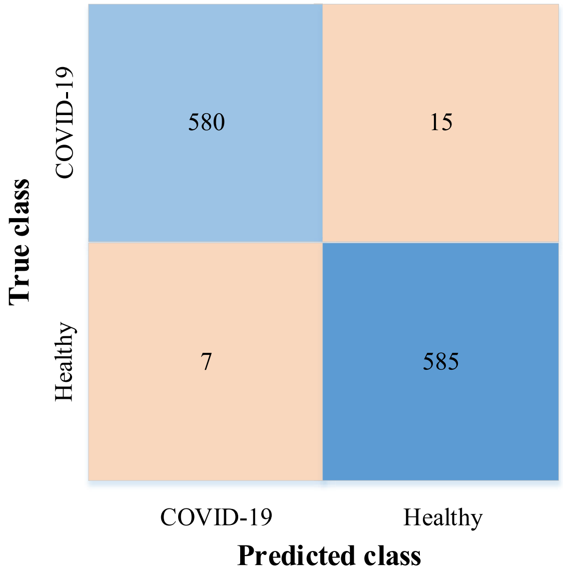
4.1. *Explainable ViT using Grad-CAM technique*

As the obtained results show that the achievements of the proposed method on all datasets were quite good, it is worth visualizing which region of the input images were efficient in detecting the COVID-19, healthy, and symptomatic classes by using the ViT model. To this end, the well-known Grad-CAM approach was considered [42]. Grad-CAM enables the viewing of every model layer and the examination of each feature map layer, both of which are required for understanding how input values influence model categorization. For example, Jahmunah et al. [43] used the Grad-Cam method to develop an explainable deep learning model to detect myocardial infarction from ECG signals. In this work, we also used Grad-CAM on the output of the ViT model. Fig. 14 depicts the activation maps generated by Grad-CAM overlaid on the FD images for the COVID-19, healthy and symptomatic classes, respectively. It is worth mention that in Grad-Cam



True class \ Predicted class	COVID-19	Healthy	Symptomatic
COVID-19	614	25	0
Healthy	20	6500	44
Symptomatic	0	49	1642

Fig. 8 – Cumulative confusion matrix obtained for the COUGHVID dataset.



True class \ Predicted class	COVID-19	Healthy
COVID-19	580	15
Healthy	7	585

Fig. 10 – Cumulative confusion matrix obtained for the VIRUFY dataset.

images, while the red color indicates the most relevant region during the classification process, the blue color indicates the less relevant region during the classification process. As seen in Fig. 14, for the COVID-19 class, the ViT model generally concentrated on the Castigloni and Higuchi FD regions. Similarly, for the healthy class, the ViT model gave its decision mostly based on the Castigloni FD, where the red color density is high. Finally, the proposed model generally used the Katz and Castigloni FD to decide the symptomatic class. As the FD images for COVID-19, Healthy, and Symptomatic classes are explored, it is seen that each class has its texture structure. For example, when the FD image obtained for the COVID-19 class is examined, it can produce very different and sensitive features due to the large variation between segments and, therefore, the variable time scale found in the Castigloni and Higuchi fractal calculation methods. In other words, the mathematical structures of these fractal calculation methods come to the fore more decisively in class distinctions.

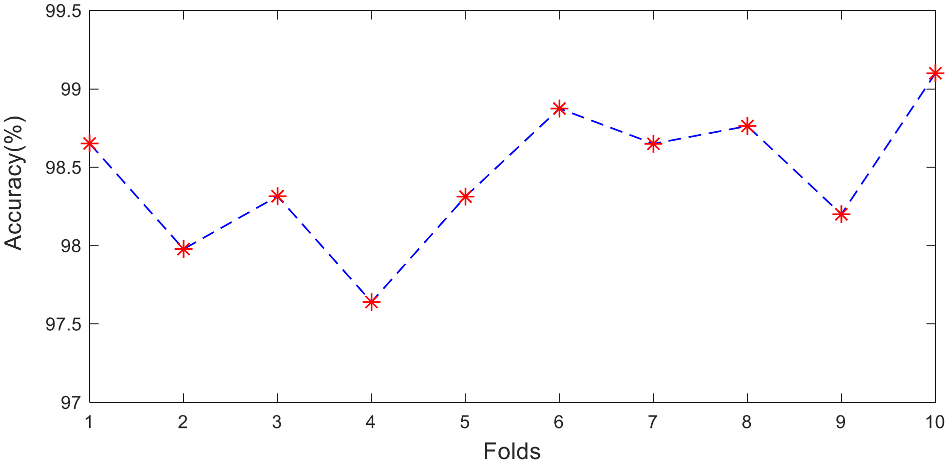


Fig. 9 – Graph of accuracy (%) obtained scores for each fold using the COUGHVID dataset.

Table 2 – Performance evaluation scores for the VIRUFY dataset.

	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)
VIRUFY	98.15	97.48	98.82	98.14

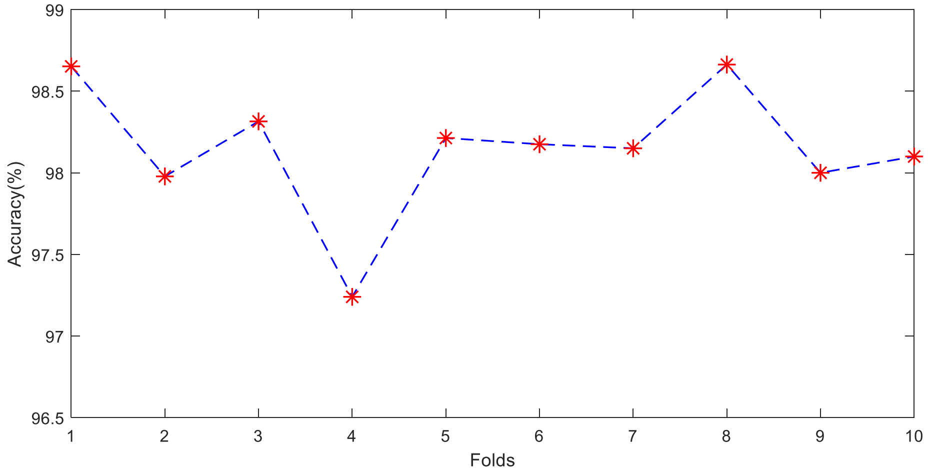


Fig. 11 – Graph of accuracy (%) obtained scores for each fold using the VIRUFY dataset.

Table 3 – Performance evaluation scores obtained for the COSWARA dataset.

	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-score (%)
COSWARA	97.59	88.44	98.73	89.04

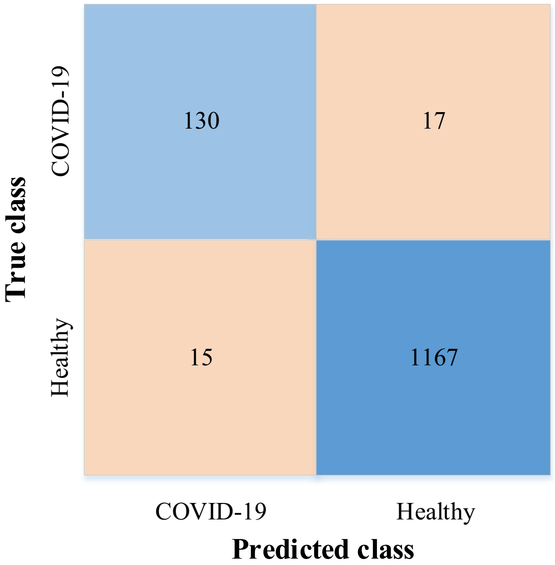


Fig. 12 – Cumulative confusion matrix obtained for the COSWARA dataset.

5. Discussions

This paper proposed a novel, explainable approach for cough sound-based COVID-19 recognition. In the proposed method, the cough sound is divided into overlapping segments, and each segment is labeled since the input audio may contain other noises. The deep YAMNet model is used to perform this. Following labeling, the cough segments are cropped and concatenated to recreate the pure cough sounds. The fractal coefficients on the cough sound are then calculated using four different fractal dimension approaches with an overlapping sliding window that generates a matrix. The fractal dimension images are then produced using the built matrixes. Finally, the produced images are classified into COVID-19, healthy, and symptomatic classes using a ViT model that has been pre-trained. As shown in Tables 1, 2, and 3, the fine-tuned ViT model with fractal images yielded 98.45 %, 98.15 %, and 97.59 % accuracy scores for COUGHVID, VIRUFY, and COSWARA datasets, respectively. As the COUGHVID dataset contains a higher number of samples, the accuracy score for this dataset was higher than the other datasets. In the recent literature, various approaches have been used for cough-based COVID-19 diagnosis. Table 4 shows a comparative summary of the approaches that were conducted on cough-based COVID-19 diagnosis. As seen in Table 4, Islam et al. [16] and Manshouri [18] used the VIRUFY dataset and obtained 93.8 % and 95.86 % accuracy scores with their proposed methods. Hamdi et al. [17] used the COUGHVID dataset and obtained a 91.13 % accuracy score. Rahman et al [21] used CAMBRIDGE and QATARI datasets and obtained a 96.5 % accuracy score. Ren et al. [22] used the COUGHVID dataset and obtained a 0.632 Unweighted Average Recall (UAR) score. Andreu-Perez et al. [25] used their own dataset with CNN and obtained 97.18 % and 96.64 % average accuracy scores for COVID-19 and healthy classes. Zealouk et al. [26] also used their dataset with HMM classifier and obtained 93.33 % and 86.66 % average accuracy scores for COVID-19 and healthy classes. Son et al. [44] used both MFCC and spectrogram images and deep feature extraction and deep neural networks (DNN) for COVID-19 detection. They obtained 88 %, 62 %, and 94 % classification accuracy scores using CNN, LSTM, and DNN, respectively. For COVID-19 cough identification, Xue

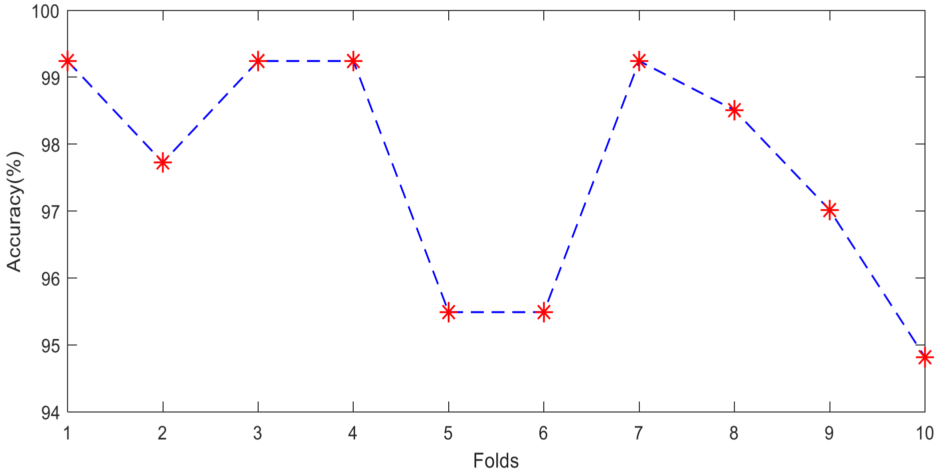


Fig. 13 – Graph of accuracy (%) obtained for each fold using the COSWARA dataset.

et al. [45] developed a self-supervised learning-enabled system. A contrastive pre-training phase was used to train a transformer-based feature encoder with unlabeled data. In deep neural networks, Soltanian et al. [46] employed a mix of quadratic kernels and the notion of separable kernels to improve recognition accuracy concurrently.

Xue et al. [45] used MFCC, log-compressed mel-filterbank features, and VGGish and Transformer-CP classifiers on the COSWARA dataset for COVID-19 identification. Authors obtained 83.15 % and 83.74 % accuracy scores for VGGish and Transformer-CP classifiers, respectively. Soltanian et al. [46] used MFCC images with ordinary CNN and a novel CNN model to detect COVID-19 on the VIRUFY dataset.

As we would like to show the robustness of the proposed approach, we also used the 5-fold cross-validation test in our work, and the related accuracy scores were given in Table 4. As seen in Table 4, 98.37 %, 97.38 %, and 98.24 % accuracy scores were obtained for COUGHVID, COSWARA, and VIRUFY datasets, respectively. The obtained results with the 5-fold cross-validation were quite similar to the results that were obtained with the 10-fold cross validation which indicated the robustness of the proposed works. In addition, authors obtained 95 % and 97.5 % accuracy scores for ordinary CNN and Separable quadratic CNN models, respectively.

As the literature about the related works were examined, it was seen that there has been various cough datasets for COVID-19 detection. We actually used the dataset that were publically available and could be used for performance comparison purposes. Authors have also opted to produce their own datasets and use them in their works. And we preferred to put them in Table 4 for comparison purposes.

The advantages of this study are given below:

1) Cough sound segments detection and concatenation removed the unwanted segments in the input audio signal. This makes the approach more robust against the noise, speech, and silence sound segments.
2) The proposed approach obtained 98.45 %, 98.15 %, and 97.59 % accuracy scores for COUGHVID, COSWARA, and VIRUFY datasets, respectively. To the best of our knowledge, this is the first study to use three public datasets and obtain the highest classification accuracy.
3) Construction of FD images with a sliding window enables investigation of all cough sounds' details, which produced more accurate and robust results.
4) In this work, we also demonstrated which region of the input image was important when the ViT model was given a decision about the class labels. Thus, an explainable model was represented instead of a black box representation.

The limitation of our method is as follows:

1) Fine-tuning the ViT model is time-consuming when 5 and 10-fold cross cross-validations are employed.

6. Conclusions

This paper proposed a novel and explainable approach for accurate and contactless COVID-19 detection using cough sounds. We save the proposed method not to be a black box model by introducing explainability. The proposed approach uses a preprocessing stage for cough sound segmentation from the input audio signal using YAMNet architecture. The segmented cough sound segments were then concatenated to form the final input cough sounds. Four FD approaches were employed to convert the cough sound into FD coefficients, and these coefficients were located in the rows of a matrix. The constructed matrix was then saved as image. These images were then fed to the pretrained ViT model to train the pretrained ViT model further. In this work, we obtained an accuracy score of 98.45 %, 98.15 %, and 97.59 % for the COUGHVID, VIRUFY, and COSWARA datasets, respectively. The main limitation of this work was that the complex-

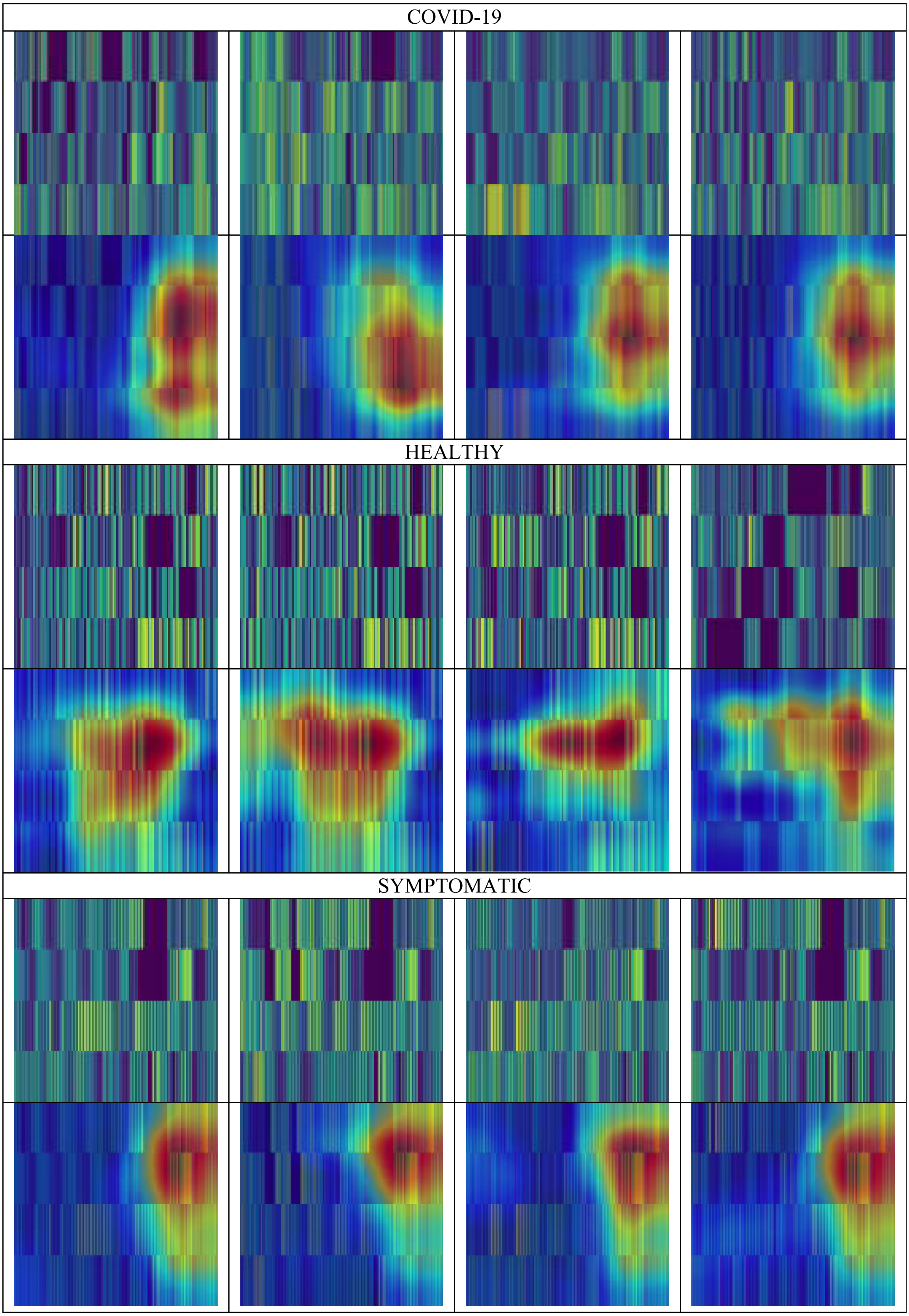


Fig. 14 – Heat maps of each cough sound class obtained from the Grad-CAM technique.

Table 4 – Comparison of the proposed approach with existing approaches.

Study	Year	Dataset	Extracted features	Classifier	Validation Method	Accuracy
Islam et al. [16]	2022	VIRUFY	Mixture of time and frequency domain features	DNN	Cross-validation (5-fold)	93.8 %
Hamdi et al. [17]	2022	COUGVID	Spectral features	CNN + LSTM	Cross-validation (10-fold)	91.13 %
Manshouri [18]	2022	VIRUFY	Spectral features	SVM	Cross-validation (LOO)	95.86 %
Rahman et al. [21]	2022	CAMBRIDGE and QATARI	Stacking of deep features	Logistic regression (LR)	Cross-validation (5-fold)	96.5 %
Ren et al. [22]	2022	COUGVID	Acoustic features	SVM	Cross-validation (5-fold)	0.632 (UAR)
Andreu-Perez et al. [25]	2022	Author's own dataset	Audio sonography	CNN	Cross-validation (10-fold)	97.18 % COVID-19, 96.64 % Healthy
Zealouk et al. [26]	2022	Author's own dataset	MFCC	HMM	Cross-validation (LOO)	93.33 % COVID-19, 86.66 % Healthy
Son et al. [44]	2022	COUGHVID	MFCC	LSTM	Hold out (70:15:15)	62 %
Son et al. [44]	2022	COUGHVID	Spectrogram and deep features	DNN	Hold out (70:15:15)	94 %
Xue et al. [45]	2021	COSWARA	MFCC + log compressed mel-filterbank	VGGish	Hold out (70:10:20)	83.15 %
Xue et al. [45]	2021	COSWARA	MFCC + log compressed mel-filterbank	Transformer-CP	Hold out (70:10:20)	83.74 %
Soltanian et al. [46]	2022	VIRUFY	MFCC	CNN	Cross-validation (5-fold)	95 %
Soltanian et al. [46]	2022	VIRUFY	MFCC	Separable quadratic CNN	Cross-validation (5-fold)	97.5 %
Proposed	2022	COUGVID	FD images	ViT	Cross-validation (10- fold, 5-fold)	98.45 % 98.37 %
Proposed	2022	COSWARA	FD images	ViT	Cross-validation (10- fold, 5- fold)	97.59 % 97.38 %
Proposed	2022	VIRUFY	FD images	ViT	Cross-validation (10- fold, 5- fold)	98.15 % 98.24 %

ity of fine-tuning the ViT model was too high, with a 10-fold and 5-fold cross-validation strategy. In addition, the performance of the YAMNet model was quite important for the subsequent processes of the proposed work. If YAMNet missed a cough sound segment, that segment may not be further processed. In the future, we plan to use this proposed model to detect other respiratory disorders like asthma, COPD (chronic obstructive pulmonary disorder), and bronchitis.

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.

REFERENCES

[1] Pandemic 2020 emergencies/diseases/novel-coronavirus-2019 https://www.who.int/.

[2] Sengur D. Investigation of the relationships of the students' academic level and gender with Covid-19 based anxiety and protective behaviors: A data mining approach. Turkish J. Sci. Technol. 2020;15(2):93–9.

[3] Lan L, Xu D, Ye G, Xia C, Wang S, Li Y, et al. Positive RT-PCR test results in patients recovered from COVID-19. Jama Network 2020;323(15):1502–3.

[4] Alqudaihi KS, Aslam N, Khan IU, Almuhaideb AM, Alsunaidi SJ, Ibrahim NMAR, et al. Cough sound detection and diagnosis using artificial intelligence techniques: challenges and opportunities. IEEE Access 2021;9:102327–44.

[5] Pahar M, Klopper M, Warren R, Niesler T. COVID-19 Cough Classification using machine learning and global smartphone recordings. Comput Biol Med 2021;135(104572):1–10.

[6] Laguarta J, Hueto F, Subirana B. COVID-19 artificial intelligence diagnosis using only cough recordings. IEEE Open J Eng Med Biol 2020;1:275–81.

[7] Alsabek M B, Shahin I, Hassan A. Studying the similarity of COVID-19 sounds based on correlation analysis of MFCC. In: International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI) 2008 November 3 (pp. 1-5). IEEE.

[8] Sharma N, Krishnan P, Kumar R, Ramoji S, et al. Coswara-A database of breathing, cough, and voice sounds for COVID-19 diagnosis. In: Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech 2020 October 25 (pp. 4811-4815).ISCA.

[9] Mouawad P, Dubnov T, Dubnov S. Robust detection of COVID-19 in cough sounds. SN Computer Science 2021;2(34):1–13.

[10] Erdoğan YE, Narin A. COVID-19 detection with traditional and deep features on cough acoustic signals. Comput Biol Med 2021;136(104765):1–10.

[11] Despotovic V, Ismael M, Cornil M, Mc Call R, et al. Detection of COVID-19 from voice, cough and breathing patterns: dataset and preliminary results. Comput Biol Med 2021;138 (104944):1–9.

[12] Tena A, Clarià F, Solsona F. Automated detection of COVID-19 cough. Biomed. Signal Process. Control 2022;71(103175):1–11.

[13] Kobat MA, Kivrak T, Barua PD, Tuncer T, et al. Automated COVID-19 and heart failure detection using DNA pattern technique with cough sounds. Diagnostics 1962;2021 (11):1–15.

[14] Chang Y, Jing X, Ren Z, Schuller BW. CovNet: A transfer learning framework for automatic COVID-19 detection from crowd-sourced cough sounds. Frontiers in Digital Health 2021;3(799067):1–11.

[15] Chowdhury NK, Kabir MA, Rahman MM, Islam SMS. Machine learning for detecting COVID-19 from cough sounds: An ensemble-based MCDM method. Comput Biol Med 2022;145 (105405):1–14.

[16] Islam R, Abdel-Raheem E, Tarique M. A study of using cough sounds and deep neural networks for the early detection of COVID-19. Biomed Eng Adv 2022;3(100025):1–13.

[17] Hamdi S, Oussalah M, Moussaoui A, Saidi M. Attention-based hybrid CNN-LSTM and spectral data augmentation for COVID-19 diagnosis from cough sound. J Intellig Informat Syst 2022:1–23.

[18] Manshouri NM. Identifying COVID-19 by using spectral analysis of cough recordings: a distinctive classification study. Cogn Neurodyn 2022;16(1):239–53.

[19] Lee GT, Nam H, Kim SH, Choi SM, Kim Y, Park YH. Deep learning based cough detection camera using enhanced features. Expert Syst Appl 2022;206(117811):1–20.

[20] Dang T, Han J, Xia T, Spathis D, Bondareva E, et al. Exploring longitudinal cough, breath, and voice data for COVID-19 progression prediction via sequential deep learning: model development and validation. J Med Intern Res 2022: 24(6), e37004; 1-14.

[21] Rahman T, Ibtehaz N, Khandakar A, Hossain MSA, Mekki YMS, Ezeddin M, et al. QUCoughScope: An intelligent application to Detect COVID-19 patients using cough and breath sounds. Diagnostics 2022;12(4), 920:1–12.

[22] Ren Z, Chang Y, Bartl-Pokorny KD, Pokorny FB, Schuller BW. The acoustic dissection of cough: diving into machine listening-based COVID-19 analysis and detection. J. Voice 2022;36(6):1–14.

[23] Sharan P. Automated discrimination of cough in audio recordings: A scoping review. Frontiers in Signal Processing 2022;2(759684):1–18.

[24] Gabaldón-Figueira JC, Keen E, Giménez G, Orrillo V, Blavia I, et al. Acoustic surveillance of cough for detecting respiratory disease using artificial intelligence. ERJ Open Research 2022;8 (2):1–9.

[25] Andreu-Perez J, Perez-Espinosa H, Timonet E, Kiani M, Girón-Pérez MI, et al. A generic deep learning based cough analysis system from clinically validated samples for point-of-need COVID-19 test and severity levels. IEEE Trans Serv Comput 2021;15(3):1220–32.

[26] Zealouk O, Satori H, Hamidi M, Laaidi N, Salek A, Satori K. Analysis of COVID-19 resulting cough using formants and automatic speech recognition system. J. Voice 2021;36(5):1–8.

[27] YAMNet, YAMNet neural network, 2021 https://github.com/tensorflow/models/tree/master/research/ audioset/yamnet.

[28] Atila O, Şengür A. Attention guided 3D CNN-LSTM model for accurate speech based emotion recognition. Appl Acoust 2021;182(108260):1–11.

[29] Howard AG, Menglong Z, Chen B, Kalenichenko D, Wang W, Weyand T, et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. Cornell University arXiv. Comput Vis Patt Recogn 2017;1704:1–9.

[30] Gemmeke J F, Ellis D P W, Freedman D, Jansen A, Lawrence W, et al. Audio Set: An ontology and human-labeled dataset for audio events. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2017 March 5 (pp. 776-780).IEEE.

[31] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Llion Jones, Gomez A N, et al. Attention is all you need. In: 31st Conference on Neural Information Processing Systems (NIPS) 2017 December 4 (pp. 1-15). Curran Associates Inc.

[32] Katz AJ, Thompson AH. Fractal sandstone pores: implications for conductivity and pore formation. Phys Rev Lett 1985;54 (12):1325–8.

[33] Higuchi T. Approach to an irregular time series on the basis of the fractal theory. Phys D: Nonlinear Phenom 1988;31 (2):277–83.

[34] Petrosian A. Kolmogorov complexity of finite sequences and recognition of different preictal EEG patterns. In: Proceedings Eighth IEEE Symposium on Computer-Based Medical Systems 1995 June 9 (pp. 212-217). IEEE.

[35] Castiglioni P. Letter to the Editor: What is wrong in Katz's method? Comments on:" A note on fractal dimensions of biomedical waveforms. Comput Biol Med 2010;40(11–12):950–2.

[36] Orlandic L, Teijeiro T, Atienza D. The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. Sci Data 2021;8:1–10.

[37] Chaudhari G, Jiang X, Fakhry A, Han A, Xiao J, et al. Virufy: Global applicability of crowdsourced and clinical datasets for AI detection of COVID-19 from cough. Cornell University arXiv, Sound 2011;13320:1–8.

[38] Sharma N, Krishnan P, Kumar R, Ramoji S, Chetupalli S R, et al. Coswara-A database of breathing, cough, and voice sounds for COVID-19 diagnosis. In: Proceedings Interspeech 2020 October 25 (pp. 4811-4815). Interspeech.

[39] Hugging Face, Fine-Tune ViT for image classification with transformers, 2021, https://huggingface.co /blog/fine-tune-vit.

[40] Deniz E, Şengür A, Kadiroğlu Z, Guo Y, et al. Transfer learning based histopathologic image classification for breast cancer detection. Health Informat Sci Syst 2018;6(18):1–7.

[41] Kadiroğlu Z, Şengür A, Deniz E. Classification of histopathological breast cancer images with low level texture features. In: International Engineering and Natural Sciences Conference (IENSC) 2018, November 14 (pp. 1765-1772). INESEG.

[42] Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-Cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision (ICCV) 2017 October 22 (pp. 618-626). IEEE.

[43] Jahmunah V, Ng EYK, Tan RS, Oh SL, Acharya UR. Explainable detection of myocardial infarction using deep learning models with Grad-CAM technique on ECG signals. Comput Biol Med 2022;146(105550):1–19.

[44] Son MJ, Lee SP. COVID-19 diagnosis from crowdsourced cough sound data. Appl Sci 2022;12(4):1–12.

[45] Xue H, Salim F D. Exploring self-supervised representation ensembles for covid-19 cough classification. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining 2021 August 14 (pp. 1944-1952). KDD.

[46] Soltanian M, Borna K. Covid-19 recognition from cough sounds using lightweight separable-quadratic convolutional network. Biomed Signal Process Control 2022;72 (103333):1–10.