Computers, Materials & Continua
DOI: 10.32604/cmc.2022.023830
Article

Tech Science Press

Multi-Modality and Feature Fusion-Based COVID-19 Detection Through Long Short-Term Memory

Noureen Fatima[1], Rashid Jahangir[2], Ghulam Mujtaba[1], Adnan Akhunzada[3,*],
Zahid Hussain Shaikh[4] and Faiza Qureshi[1]

[1]Center of Excellence for Robotics, Artificial Intelligence, and Blockchain, Department of Computer Science, Sukkur IBA University, Sukkur, Pakistan
[2]Department of Computer Science, COMSATS University Islamabad–Vehari Campus, Pakistan
[3]Faculty of Computing and Informatics, University Malaysia Sabah, Kota Kinabalu, 88400, Malaysia
[4]Department of Mathematics, Sukkur IBA University, Sukkur, Pakistan
*Corresponding Author: Adnan Akhunzada. Email: adnan.akhunzada@ums.edu.my

Received: 23 September 2021; Accepted: 29 December 2021

Abstract: The Coronavirus Disease 2019 (COVID-19) pandemic poses the worldwide challenges surpassing the boundaries of country, religion, race, and economy. The current benchmark method for the detection of COVID-19 is the reverse transcription polymerase chain reaction (RT-PCR) testing. Nevertheless, this testing method is accurate enough for the diagnosis of COVID-19. However, it is time-consuming, expensive, expert-dependent, and violates social distancing. In this paper, this research proposed an effective multi-modality-based and feature fusion-based (MMFF) COVID-19 detection technique through deep neural networks. In multi-modality, we have utilized the cough samples, breathe samples and sound samples of healthy as well as COVID-19 patients from publicly available COSWARA dataset. Extensive set of experimental analyses were performed to evaluate the performance of our proposed approach. Several useful features were extracted from the aforementioned modalities that were then fed as an input to long short-term memory recurrent neural network algorithms for the classification purpose. Extensive set of experimental analyses were performed to evaluate the performance of our proposed approach. The experimental results showed that our proposed approach outperformed compared to four baseline approaches published recently. We believe that our proposed technique will assists potential users to diagnose the COVID-19 without the intervention of any expert in minimum amount of time.

Keywords: Covid-19 detection; long short-term memory; feature fusion; deep learning; audio classification

1 Introduction

The novel COVID-19 has been declared as a pandemic by World Health Organization (WHO) because of its rapid spread. Its trajectory growth began on January 4, 2020, which constrained most of the countries to take serious precautionary measures, such as lockdowns and dedicated isolation facilities in hospitals, to keep the infection rate at a minimum. As of mid-January 2021, the number of confirmed cases of COVID-19 has crossed 95 million with more than two million deaths. Because of its devastation, COVID-19 has put millions of lives at stake in 221 countries and territories. The global effort to address the challenge has empowered two leading companies: Moderna and Pfizer-BioNTech, which developed a vaccine for the disease with reported efficacy exceeding 90% [1]. Furthermore, dozens of vaccines are under clinical trials and around 20 are in their last stages of testing 3.

One of the most effective methods to control the spread of COVID-19 is self-isolation. The isolation period of COVID-19 may take two weeks on average [2]. The most prominent symptom found in COVID-19 patients is the failure of the respiratory system in the guise of dry cough and dyspnea; more severe condition causes rhinorrhea and sore throat SARS-CoV positive patients, after 7-10 days of infection, may show unconventional radiographic variations in lungs, thereby indicating pneumonia. About 70% to 90% of CoV-positive patients may suffer from Lymphocytopenia [3].

Real-time PCR is the most practiced method for quantifying the unique sequence of viruses in the designated gene, Ribonucleic Acid (RNA), with results available in 2–48 h [4]. This method, though generally employed to diagnose COVID-19, is inadequate to control the disease for certain reasons: a) dearth of skillful paramedical [5] At times when RT-PCR diagnosis detects COVID-19 in a patient, the virus is already spread.

Corona cases have increased with such rapidity that their increase has brought about an outgrowth in proposals on technological resolutions for healthcare. Certainly, the need for the development of modest, economical, fast, and accurate testing procedures for COVID-19 diagnosis has become pivotal for healthcare, policymaking, and economic growth in several nations. The main focus of this study is to use machine learning-based and/or deep learning-based techniques to provide an efficient model for diagnosis of COVID-19 as an alternative to traditional and cheaper alternative of RT-PCR test.

Audio signals generated by the human body (e.g., sighs, breathing, heart, digestion, vibration sounds) have often been used by clinical research to diagnose diseases. Researchers have used the human voice to diagnose the earlier symptoms of diseases such as ParkinsonâĂŹs disease correlates with the softness of speech, vocal tone, pitch, rhythm, rate, and volume correlate with invisible illnesses such as post-traumatic stress disorder [6].

Deep Learning (DL) is an area of Artificial intelligence that enables the creation of end-to-end models to achieve promised results using input data, without the need for manual feature extraction [7–9]. The best approach to these models is these models learn rich features from given raw data instead of human-engineered features. The deep learning models work effectively due to the multiple layered approaches and the model extracts more features than a human-engineered feature. Deep learning techniques have been successfully applied in many problems. For instance, arrhythmia detection [4,10], skin cancer classification [11], breast cancer detection [12]. brain disease classification, image segmentation [13] many others.

Several researchers have employed machine learning and DL techniques to detect COVID-19 through various modalities including, X-ray images [14], patient sound, breathing, and cough sound [15–19]. For instance, Imran, et al. [17] employed machine learning-based and deep learning-based approaches to identify COVID-19 patients using cough modality and achieved 93% classification accuracy. Bagad, et al. [15] applied a DL-based approach to identify COVID-19 from cough sound and yielded 90% sensitivity. In the aforementioned studies, authors have used only one modality to detect COVID-19 and their obtained accuracies can be further improved. Furthermore, in several studies, it has been reported that the combination of multi-modalities features and their fusion can further generate robust results [20,21]. To facilitate this, Sharma, et al. [18] provided an open-access dataset named COWSARA for the detection of COVID-19. This dataset includes several modalities such as breathing sound, patient sound, cough sound, and some other features like smoker, temperature, etc. Sharma, et al. [18] used the COWSARA dataset to detect COVID-19 patients. The authors achieved 63% accuracy using the random forest algorithm. However, this accuracy was low and can be enhanced with the use of several new ML-based algorithms. Thus, to overcome this, Grant, et al. [22] used the same COWSARA dataset and obtained 87% accuracy. There are two major limitations in the work done by Bagad, et al. [15] and Imran, et al. [17]. Firstly, the authors used only one modality from the COWSARA dataset that was the patient sound modality. Secondly, the accuracy is low and can further be improved with the help of more features from multiple modalities and with the help of employing more robust deep neural network algorithms Many researchers have proven that a single modality in the field of medicine is sometimes ineffective to differentiate complex detail of any disease [23] and highly suggested using the multi-modality for a better result. Keeping given those suggestions, we have proposed MMFFT.

To address this issue, we proposed a multi-modality feature fusion-based (MMFF) technique through a deep neural network for the classification of COVID-19 patients using the COSWARA dataset. Therefore, we proved our hypothesis by utilizing the cough samples, breathe samples, and sound samples of healthy as well as COVID-19. Afterward, several useful features were extracted from the aforementioned modalities that were then fed as an input to long short term memory recurrent neural network algorithm for the classification purpose. An extensive set of experimental analyses were performed to evaluate the performance of our proposed approach. The experimental results showed that our proposed approach outperformed compared to four baseline approaches published recently.

This research aims to develop an effective multi-modality-based and feature fusion-based COVID-19 detection technique through deep neural networks. In multi-modality, we aim to use the cough samples, breathe samples, and sound samples of healthy as well as COVID-19 patients from the publicly available COSWARA dataset.

The purpose of this purposed study is to answer the following questions:

RQ1: How MMFFT is beneficial for diagnosis of COVID-19 from sound, cough and breathe samples?

RQ2: What dataset has been utilized for diagnosis of COVID-19?

RQ3: How to deal with imbalance COSWARA dataset problem?

RQ4: What techniques are being employed to classify either the sound sample is belonging to healthy or COVID-19 patient?

RQ5: How our proposed model is performing compared to four existing baseline techniques?

We believe that our proposed methodology, "a multi-modality and feature fusion-based COVID-19 detection through Long Short-Term Memory (LSTM)" can be effective in enabling multi-modality-based technology solution for point-of-care detection of COVID-19, and shortly, this can help to detect the COVID-19. Moreover, this method provides COVID-19 detection results easily, within 2–3 min, and without violating social distance. Moreover, this research will provide new directions to researchers who will pursue research on COVID detection.

The rest of this study is arranged as follows. Section 2 describes existing works on COVID-19 detection techniques using machine learning or deep learning algorithms. Section 3 presents proposed techniques and proposed research methodology. Section 4 presents the experimental results. Section 5 presents the theoretical analysis of our obtained results in the form of a discussion section. Finally, Section 6 concludes our research.

2 Related Work

In this section, we discuss the algorithms and methods developed by researchers for COVID-19 detection. Several researchers have employed machine learning and deep learning to detect the COVID-19 from various modalities including, medical images and audio signals.

Image Bases Literature: Immense work has been done on radiology images for infectious COVD-19 disease diagnosis using artificial intelligence techniques. A study by [24] proposed a novel CVOIDX-Net framework for automatic identification or confirmation of COVID-19 from X-ray images using seven different models of Convolutional Neural Network(CNN); namely DenseNet121, VGG19, ResNetV2, Xception, MobileNetV2, InceptionV3, and InceptionResNetV2. The experimental results revealed that the proposed COVIDX-Net achieved the best performance on DenseNet121 and VGG19 DL models. Similarly, [14] introduced COVID-Net based on deep CNN for the COVID-19 detection from open source and publicly available CXR images. This study also investigated the explain ability method to analyze the critical factors related to COVID patients to help the clinicians. The proposed COVID-Net framework was evaluated on the COVID test data and obtained 92.4% accuracy for a DL model.

To handle the availability issue of COVID-19 test kits, [25] proposed a quick alternative for the diagnosis of COVID-19 by implementing an automatic detection system. The authors employed five pre-trained CNNs models (InceptionV3, ResNet50, ResNet101, ResNet152 and Inception-ResNetV2) to detect the COVID-19 infected patients using X-ray radiographs. The model was implemented with four classes (normal, bacterial pneumonia, viral pneumonia, and COVID-19) by using 5-fold cross-validation. The performance results show that the ResNet50 model outperformed the other four models by achieving 96.1%, 99.5%, and 99.7% accuracy on three different datasets.

Audio Signals Literature-Besides using X-ray images, various researchers have been working on the use of respiratory sound (audio) to diagnose illness and recognize sound patterns for years. COVID-19 detection using deep learning has achieved excellent results in medical images. However, the biggest challenge using Medical images modalities is that if a patient has a positive result for the RT-PCR test, at the same time of admission he must have a chest infection because chest infection may occur after one or two days of onset of symptoms [26].

Brief summary of audio-based literature is shown in Tab. 1. The authors have worked on multi-modalities to classify the COVID-19 and healthy patients [20,22,24,27] still there are some limitations. Firstly, the authors have not deal with feature fusion techniques to make a model generalized. The generalized model is more reliable, stable, and accurate while dealing with feature fusion techniques. Secondly, the authors [22,24] have used Random Forest (RF) classifier and achieved 66.74% and 87.5%. However, The Random Forest (RF) classifier is slower than the LSTM and it can be a slowdown performance of classification when RF has a large number of a tree. In previous studies, the author [16] has used CNN model on Crowdsourced dataset on two modality such as cough and breathe and achieved AUC of 80.46%. Moreover, by using the same dataset, the author [28] applied VGGish algorithms and obtained 80% (Area under Curve) AUC respectively. In addition to this, the authors have not deal with dataset imbalance [17,22,24].

Table 1: Audio based literature

Studies	Dataset	Modality	Classifier	ACC	AUC	Weakness
[17]	ESC-50 dataset	Cough	CNN	93%		Single modality and imbalanced class problem was not resolved
[15]	FreeSound database flusense COSWARA	Cough	CNN	72%		Single modality, imbalanced class problem was not resolved and, low accuracy.
[22]	COSWARA	Cough, sound and breath	RF	69%		Imbalanced class problem was not resolved and, low accuracy.
[24]	COSWARA	Cough, sound and breath	RF	87%		Imbalanced class problem was not resolved and, low accuracy.
[16]	Crowdsourced	Cough and breathe	CNN		80.46%	Low accuracy
[28]	Crowdsourced	Cough and breathe	VGISH		80%	Low accuracy and Imbalanced class problem was not resolved

3 Proposed MMFF Technique

This section presents in detail the philosophy behind the construction of our proposed MMFF technique. To summarize, for the construction of the MMFF technique, a publicly available multi-modality dataset was used for the detection of COVID-19. Afterward, several speech preprocessing techniques were employed to remove non-discriminative information from the audio signals. Subsequently, several discriminative features were extracted from each preprocessed modality to generate a master feature vector (MFV). Finally, the generated MFV was then fed as an input to the Long Short-Term Memory (LSTM) recurrent neural network algorithm to construct the COVID-19 classification model. The dataset which we used for experiments was imbalanced. Thus, audio augmentation techniques were employed to overcome the class imbalance problem. The details are available in a subsequent section.

3.1 Data Collection

COSWARA is publicly available dataset [18], realized by Indian Institute of Science (IISc) Bangalore. This dataset uses the audio files of respiratory, cough, and speech sounds of normal as well as COVID-19 patients. The samples were collected from all the regions of world except Africa as shown in Fig. 1. Voice alternation was recorded from these three modalities: cough sounds (shallow and deep), breathing sounds (slow and fast), vowel sounds (a, e, and o), counting numbers from 1 to 20 (at a slow and fast pace). There are 1400 patients' data: 97 records belong to COVID-19 patients while the rest belong to healthy patients. All the sounds have uniformed sampling rate at 44.1 KHz. Moreover, some audio files contain noise. Besides, these sound modalities, the COSWAR dataset also contains 26 more features of healthy as well as COVID-19 patients. These features include, current status of health of the patient, age, country, smoker or non-smoker, and others. Fig. 1 shows the subjectivity analysis of each age group along with female and male gender. 1 indicate the age group of 21, 2 for 26–45 age group, 3 represent for 46 to 65 age group and 4 represent the above 65 group of the age. Additionally, it also shows the percentage of each item calculated in relation to each category.

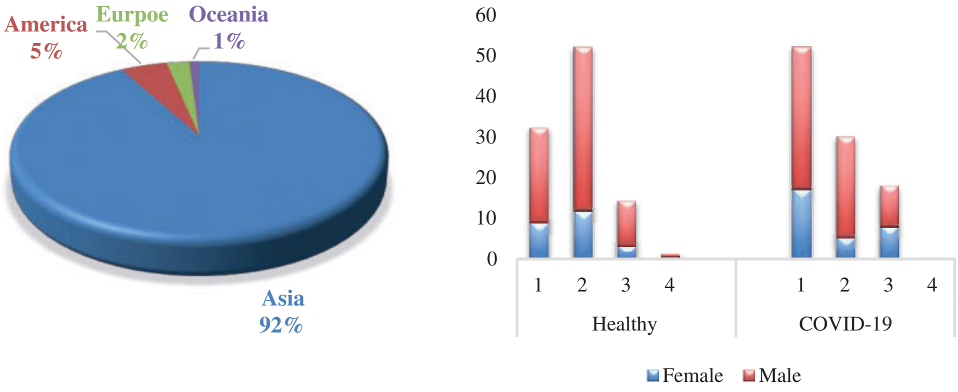

Figure 1: Pie chart of country wise classification and subjectivity analysis gender-wise

3.2 Proposed Master Features Vector

This section discusses the detail of feature fusion vector composition and Time-domain feature and frequency-based feature extraction. As shown in Algorithm 1, the master feature vector combined all five features set into a single set named it master feature vector. In the Coswara dataset, we have a folder of each patient with its P_ID. Each patient folder contains 9 types of audio sounds: cough deep, cough shallow, vowel a, vowel e, vowel o, counting fast. MFV is the master feature vector for each of the nine audios of a patient. These five features are steps are discussed in the following paragraphs.

Zero-Crossing Rate (ZCR) features were initially extracted from patient audio. Zero crossing is a rate of measure the occurrence of zero in a given time frame. The rate of significant changes along with signal for example the signal change from positive to zero or positive to negative or negative to positive. Zero crossing is an important feature in audio processing because it helping in to differentiate the percussive and pitched audio signal. Percussive audio has a random zero-crossing rate across buffer whereas pitched have a more constant value.

Root Mean Square (RMS) features were extracted from patient audio. The RMS is the root mean square of a short-time Fourier transform that provides loudness of the audio. This feature is important that will give information about a rough idea about the loudness of an audio signal.

Spectral Centroid (SCD) indicates where the center of mass of spectrum is located, and it is calculated as the weighted mean of frequency present in the signals. If the frequencies of an audio signal are the same in each number of frames, then spectral centroid must be around at the center and if there are high frequencies at the end of audio signals then spectral centroid would be at its end.

Spectral Roll-Off (SRO) indicates where the frequency lies below the specified percentage. If the frequency is cut off from the corner in dB per octave. An octave is a double of frequency. The spectral roll-off is used to differentiate the harmonic and noisy audio.

Mel Frequency Cepstral Coefficients (MFCC) We humans can't linearly interpret a pitch; various scales of frequencies are formulated to denote the way humans hear the distances between pitches. Mel scale is one of them and Mel is a unit of measure based on human perceived frequency. MFFCs are comprised of the following sub-processes: framing the audio sound, windowing, Discrete Fourier transform (DFT), the logarithm of magnitude, after that wrapping frequency on the Mel scale in the last discrete cosine transform (DCT).

3.3 Multi-Modality Fusion

Varies researchers have employed machine learning and DL techniques to detect COVID-19 through single modality including, X-ray images, patient sound, breathing, and cough sound. However, there are no prominent symptoms found in COVID-19 patients, it may be dry cough, breathing problem, sore throat, or fever. Therefore, we should not rely on a single modality to detect the COVID-19. For instance, a patient may have a COVID-19 positive result but it may not have a cough symptom similarly, it is not necessarily that a patient has a breathing problem and any other symptoms at the same time. Many researchers have proven that a single modality in the field of medicine is sometimes ineffective to differentiate complex detail of any disease [23] and highly suggested using the multi-modality for a better result. Keeping given those suggestions, we have proposed an Effective Multi-Modality and Feature_Fusion-Based COVID-19 Detection through LSTM. In which model we combine three modality sound, cough, and breathe.

3.4 Balanced Sampling and Data Augmentation

Data augmentation is a technique that creates new training data samples from existing training data to produce quantity and diversity in a dataset. This technique has been proven to effectively alleviate model overfitting [29]. Data augmentation not only improves the overall performance but also enhanced the data distribution invariant that leads to variance reduction [30]. Standard signal augmentation methods have been applied on the audio raw signals and best parameter selection as given by Nanni, et al. [31]. These are Gaussian noise, time stretch, pitch shift, and changing the speed. The detail of each is given below:

Gaussian Noise (GN) is added to raw audio samples in between variance of 0 and 1. Gaussian noise generates a new raw audio sample with preserving its originality of audio sample. It is very much important to choose the right hyper parameter for noise amplitude, σ is a notation of it. Large σ size makes it difficult to learn classifier and smaller size of σ difficult to disturb The parameter sizes in this proposed methodology were selected in between [0.004, 0.005] by following a uniform distribution.

Time Stretch (TS) is a technique for audio augmentation that changes the tempo and length of an audio clip without being changing the pitch. Υ is a parameter factor that is being added to actual audio to generate new audio? The value of Υ is in-between [0.18, 1.25] and selection of Υ is a bit tough because if the value of $\Upsilon > 1$ then audio signal may speed up and if $\Upsilon < 1$ then, a signal may slow down and increase the length of an original audio clip. The parameter sizes in this proposed methodology were selected in between [0.5, 0.8] by following a uniform distribution.

Pitch Shift (PS) is a technique that generates a new sound without changing the tempo of audio by shifting the pitch of wavelength n_steps. Time stretch is reciprocal of pitch shift. The values of n_steps should be in between [−4, 4]. The parameter n_steps in this proposed methodology were selected in between [−2, −4] by following a uniform distribution.

Changing Speed (CS) is similar to changing the pitch but here it stretches times series by a fixed rate. The parameter *n*_steps in this proposed methodology were selected in between [0.8, 0.5] by following a uniform distribution.

Algorithm 1: Algorithm to Construct Master Feature Vector

```
Input: A path to main folder of COSWARA dataset
Output: construction of master features vector comprising 5 features of all audios
folders ← count the total number of subfolders in COSWARA folders/P_ID
Initialization of variables: assign zero to variable i and j
while ($i$ < folders) do //Read all the subfolders in main folder
    audio-files ← count the total number of wav files in each subfolder
    while ($j$ < audio-files) do //Read each wav file one by one to extract features
        compute zero cross rate from audio j by using librosa.feature.zcr
        compute root mean square from audio j by using librosa.feature.rms
        compute spectral centroid from audio j by librosa.feature.spectral_centroid
        compute spectral roll-off from audio j by using librosa.feature.spectral_rolloff
        compute MFCC from audio file j by using librosa.feature.mfcc
        convert each computed features in to single column matrix
        concatenate all these single column with previous audio files master vector
        MFV = zero cross + root mean square + spectral centroid + spectral roll-off + MFCC
    end
end
```

Signal Speed (SS) in this augmentation technique we speed roll of by speedup factor range in between [0.8, 1.2]. We have applied roll-off on 0.8 and 1.1 audio signals.

Imbalanced class problems are found in many classification problems [27]. Class imbalance problems occur when the number of instances from one or more classes is considerably greater than another class [32]. In Coswara dataset 6% of dataset contain the COVID patients and 94% of dataset contain healthy patient. Such a large imbalanced dataset, up-sampling or down-sampling is difficult to implement because u up-sampling leads to uncertainty meanwhile down-sampling wastes a large portion of the data [33]. In the proposed methodology we alleviate the imbalanced class by using data augmentation to create diversity in a dataset and to make it balance. In this proposed methodology, we have 80 COVID-19 patient data, we apply 5 different augmentations on two parameters then it generated 2∗5∗80 = 800 new augmented similar sound.

3.5 Long Short-Term Memory

Long Short term Memory [34] architecture consists of three main gates namely, input, forget and output gates [35]. The hidden state is computed using these three gates.

$$ig = \sigma\left(Wi\left[x_g, h_{g-1}\right] + bi\right) \tag{1}$$

$$\mathtt{C}_1 = tanh\left(Wi\left[x_g, h_{g-1}\right] + bc\right) \tag{2}$$

$$f_g = \left(Wi\left[x_g, h_{g-1}\right] + b_f\right) \tag{3}$$

$$C_g = i_g * \mathtt{C}_1 + f_g C_{g-1} \tag{4}$$

$$O_g = \sigma\left(W_o\left[x_g, h_{g-1}\right] + b_o\right) \tag{5}$$

$$h_g = O_g \tanh\left(C_g\right) \tag{6}$$

where Eq. (1) represents the network input gate, Eq. (2) represents the memory cell of the network candidate, equation Eq. (3) represent shows the activation function of the forget gate, Eq. (4) represents the calculation of the final output gate's value for a new memory cell, and equation Eq. (5) and equation Eq. (6) explain the value of network's final output gate. Moreover, b shows the bias vector, W represents a weight matrix, shows the input to the memory cell at time g. Whereas i, c, f, o indicate the input, cell memory, forget, and output gates respectively as shown in Fig. 2

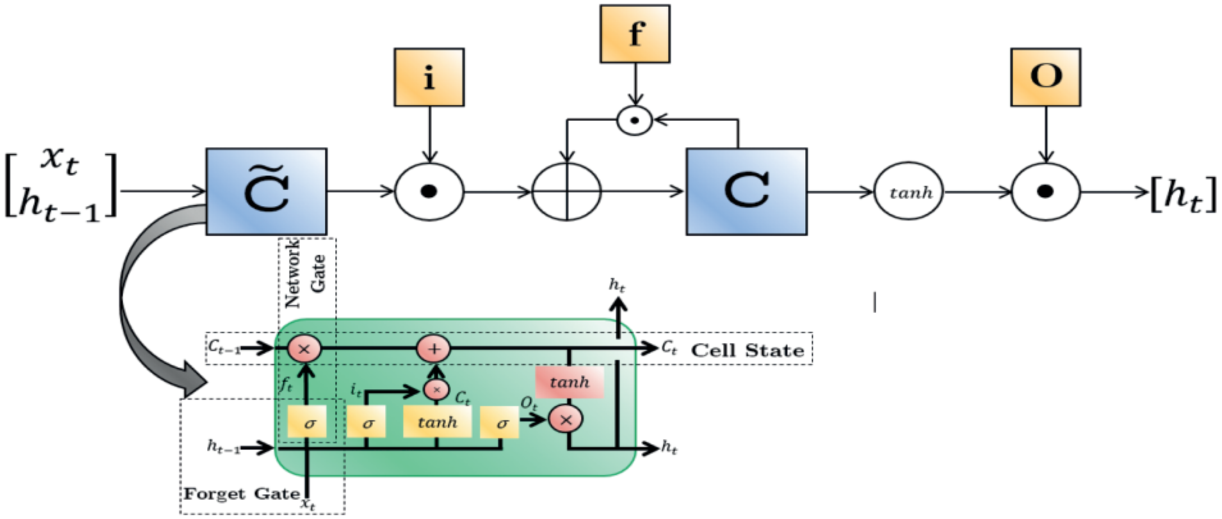

Figure 2: LSTM network

LSTM is known for learning the long-term sequences dependencies. It has the quality to learn the information for a longer period and have the decision capability to decide based on information to keep or discard. Therefore, It is proven that LSTM perform better than RNN and CNN [10]. Moreover, the LSTM network has a gated mechanism through which it controls the flow of sequences called cells. The single LSTM network cell is shown in Fig. 2. LSTM is known for its cell stage and the main purpose of the cell stage is to build the bridge for a sequence to flow the information among gates. Fig. 2, Ct represents the new cell state and Ct-1 is the old cell state. The path between these both cell sate is shown in form of horizontal for the interaction of the cells. The network gate is used to control the flow of information as the pink circle represents the multiplication operator for the pointwise and the yellow box represents the sigmoid function. Sigmoid functions work on 0 and 1, whereas 1 means need to be done and 0 means nothing needs to be done inside the network gate. Moreover, the forget gate is to check which information should be excluded and which information should include. The decision is done by checking the previous output stage. Whereas is the output of the previous stage, is the current input, and a decision is made by a sigmoid function. The input data structure consists of two neural network layers namely sigmoid and tanh. A sigmoid neural network is used to decide what value needs to be updated whereas tanh is used to create a new vector for a new candidate value. Once the new candidate has been created, it's time to entire C1. This has been done by multiplying the output of the previous stage with a new stage as shown in and equations Eq. (4) and Eq. (5). Finally, the output of a sequence is calculated via the output gate. After the sigmoid layer decides which part of the sequence has to send to the output layer, the tanh layer creates a new cell state value in between (1, −1), and the sigmoid value is multiplied with the output of selected information.

Recently, many researchers have employed LSTM for audio classification, emotion detection [10] and reported outperformance of LSTM in audio processing. The customized LSTM architecture has been designed to classify COVID-19 and non-COVID patients. The architecture consists of 1 input layer, 4 hidden layers, and 1 output layer as shown in Tab. 1. We had tried different configuration settings to opt for the best performance of LSTM. RMS, Adam, and SGD optimizers were employed with different settings of dropout and batch size. The best-fit setting was on Adam optimizer, having a 0.5 dropout value with 216 batch size. The input layer has 512 neurons, and the hidden layer consists of half of its output neurons. To remove the overfitting single dropout has been added to the network at layer 3. To overcome the overfitting problem in LSTM many researchers have successfully employed the dropout technique and successfully. The neuron section process is done carefully and effectively because a minimal number of a neuron may cause underfitting, whereas lager size of a neuron may cause overfitting [10]. Each hidden layer has Rectified Linear Unit (ReLU) function which computes output from the input within the range of 0 and 1. Finally, in the output layer, the sigmoid function has been applied and has a single neuron because the output layer neuron depends upon several classes. We applied four dropouts to overcome the overfitting issue.

3.6 Evaluation Matrices

The performance of each COVID-19 detection model was evaluated using different evaluation metrics. These evaluation metrics include accuracy (training, validation, and testing), F1-score, precision, and recall. For each sound class, the detection was measured with the labels, and the number of false-positive (FP), true positive (TP), false-negative (FN), and true negative (TN) were computed using the confusion matrix of each prediction. The mathematical representation of all the metrics are given. In addition, these evaluation metrics have been widely employed for the evaluation of various disease detection, classification and related systems [10]. ***Precision***-Precision is the ratio of correctly predicted labels for the specific class about all predicted labels of the class. It evaluates the performance of the proposed models to correctly detect actual respiratory sound. ***Recall***-Recall is the ratio of all predicted labels for the specific class to the actual labels of the class. It calculates the number of accurately detected instances as positive instances. ***Accuracy***-Compute the frequency of accurately detected respiratory sound classes from the total number of sound signals. ***F1-Score***-It computes the weighted harmonic mean/balanced ratio of recall and precision.

4 Results

This section is dedicated to the results and discussion of our experiments performed to evaluate the performance of our proposed MMFFT technique. First, we evaluated the performance of individual modalities of the COSWARA dataset, to compare the single modality experimental results with multi-modality experimental results. Secondly, our proposed MMFFT was evaluated using multi-modality and feature fusion based by using LSTM classifier algorithms. Third, we compared the performance of our proposed MMFFT technique with augmentation on raw data. In the last, we compared our purposed MMFFT with four [16,18,22,28] existing baseline techniques. The objective of this comparison was to find out the accuracy of MMFFT with the baseline. Furthermore, we perform the experiments on window 10 with GPU (GeForce MX130), using python languages.

4.1 Individual COSWARA Modality Result

In this setting, from each of nine modalities, we extracted five different types of features to prepare MFV. The constructed MFV of individual modality was then fed as an input to LSTM to evaluate its results. The results of each modality are shown in Tab. 4. As shown in the table, we have achieved accuracy in between 89% to 97% on a training set, 88% to 93% on the validation set, and 88% to 96% on the test set.

The training accuracies observed for individual modalities were in between (92% to 93%) The highest accuracy (93%) on the validation set was observed in the CHCM dataset followed by CHCM, CSCM, CFCM, and BSCM datasets (92%). The lowest validation accuracy (89%) was observed in the VACM dataset. The highest test set accuracy (94%) was observed in CHCM followed by CNCM and VECM datasets (92%). The lowest test set accuracy (88%) was observed in VFCM and VOCM. We have also reported the confusion matrix results of test set analyses in Fig. 3 to show the true positive, true negative, false positive, and false negative values of each analysis. As can be seen from Fig. 3, most of the instances were classified correctly into their respective classes. Furthermore, as can be seen from Tab. 2, the single modality models were just the right models and these were neither overfitted nor under fitted. However, the results still need further improvement in many modalities. Therefore, to further improve the results, we proposed fusion of multi-modalities in the construction of the classification model in Section 4.2.

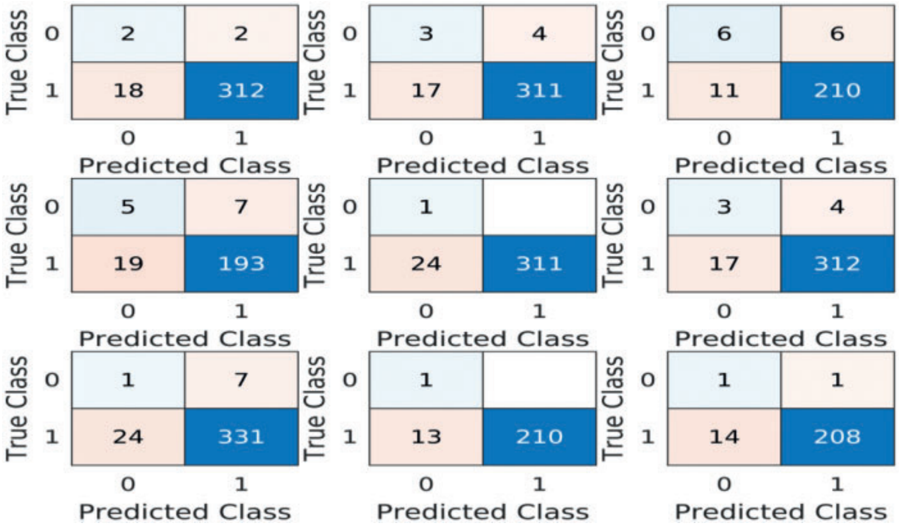

Figure 3: Confusion matrix of each modality

Table 2: LSTM configuration

Layer	Type	Neuron	Function	Dropout
1	input	512	ReLU	
2	hidden	64	ReLU	
3	hidden	32	ReLU	0.5
4	hidden	16	ReLU	
5	output	1	Sigmoid	

4.2 Multi-Modality Result

This section presents the result of multi-Modality as shown in Tab. 3. We compared the performance of our proposed MMFFT technique. In this setting, we first prepared nine MFVs (one from each modality) then we combined all nine MFVs to prepare one super MFV. Finally, we gave this super MFV as an input to the LSTM algorithm to evaluate the performance of our proposed MMFFT technique.

Table 3: Performance of combined modalities using LSTM

Augmentation	Precision-M	Recall	F1-score	Testing accuracy
Without	0.96	0.97	0.96	0.94
With	0.97	0.99	0.97	0.96

We run 100 epochs to train our proposed model by using LSTM. Nevertheless, we observed the highest training accuracy after 12 epochs and after 12 epochs the model started cramming or overfitting. The fusion of all nine modalities shows better results compared to a single modality. The experimental results of the single modality showed 94% testing accuracy and 96.5% F1-score. Fig. 4 shows the confusion matrix of our proposed COVID-19 classification for multi-modality. As can be seen from this figure, most of the instances were correctly classified into their respective classes except six instances which were healthy patients but were classified as COVID-19 patients by our proposed MMFFT-based model.

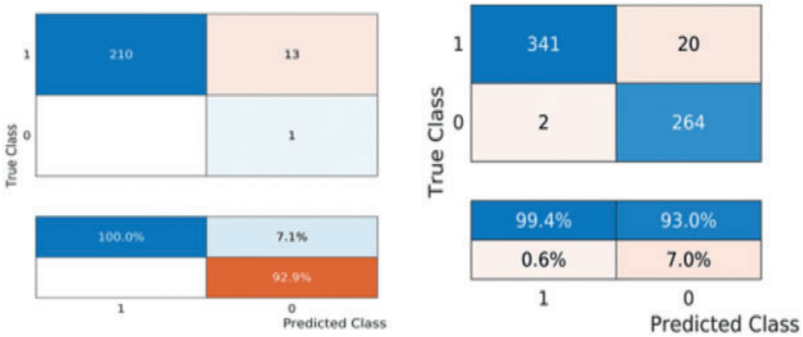

Figure 4: Confusion matrix for COVID-19 detection with and without data augmentation

4.3 Multi-Modality with Data Augmentation

This section presents the result for multi-modality with data augmentation, results are shown in Tab. 3). The COSWARA dataset is an imbalance in nature where the instances of the COVID-19 class are too much smaller in number than the healthy class. We employed five different augmentation techniques to improve the number of minority classes as discussed in Section 3.5. After performing the audio augmentation, we prepared nine MFVs (one from each modality) then we combined all nine MFVs to prepare one super MFV. Finally, we gave this super MFV as an input to LSTM to evaluate the performance of our proposed MMFFT technique with audio augmentation. We run 100 epochs to train our proposed model with audio augmentation by using LSTM. The fusion of all nine modalities with audio augmentation showed improvement of 2% of performances (94% testing accuracy) compared to experimental setting-II (96% training accuracy). Fig. 4 shows the confusion matrix of experimental setting-III. As can be seen here 20 instances were falsely positive and 2 were false negative and the rest of the instances were classified correctly into their respective classes. This shows that the audio augmentation showed marginal higher results compare to multi-modality.

Fig. 4 shows the confusion matrices of our proposed COVID-19 classification for multi-modality with augmentation. In this setting proposed methodology, for COVID-19 class, 264 instances were correctly classified whereas 2 instances were misclassified. For the non-COVID class, 341 instances were classified correctly and 20 were incorrect. The main reason for not being classified correctly is maybe data is similar or sequences almost resemble each other.

4.4 Comparing the Result of Proposed MMFFT with Baselines Methods

To show the effectiveness of our proposed COVID-19 Classification coupled with LSTM. We compared the performance of our proposed MMFFT technique with four baseline techniques recently published in the literature. The detailed results are shown in Tab. 4. The author's experimental results showed 66.74% accuracy on the test set. The results were further improved [22]. In [22] authors applied feature fusion and Random Forest classifier and obtained 87.5 & classification accuracy. In [16] and [10] authors used a Crowdsourced dataset where they used cough and breathe modalities and fed the features of these modalities to CNN and VGGish algorithms and obtained 80.46% and 80% AUC respectively. It can be seen from Tab. 4, our proposed MMFFT technique by using LSTM outperformed and showed 17% improved accuracy compared to baseline techniques

Table 4: Performance of each modality with feature fusion and LSTM

Modality	Accuracy		Precision	Recall	F1-Score
	Validation	Testing			
CHCM	0.91	0.94	0.94	0.99	0.96
CSCM	0.92	0.91	0.93	0.97	0.95
CNCM	0.92	0.92	0.93	0.99	0.96
VACM	0.89	0.90	0.93	0.96	0.94
VECM	0.91	0.88	0.92	0.95	0.94
VOCM	0.91	0.88	0.93	0.93	0.93
CFCM	0.93	0.92	0.94	0.98	0.96
BDCM	0.93	0.89	0.93	0.96	0.94
BSCM	0.91	0.92	0.94	0.98	0.96

5 Discussion

This section presents the analysis and significant result of the proposed MMFFT classification model using the COSWRA dataset. The proposed MMFFT technique obtained reliable and improved classification performance. The rigorous experimental evaluation of complex, challenging, and standard publicly available COSWRA dataset proved that the MMFFT technique is less complex, accurate, reliable, and more effective than the existing baseline techniques. Several studies have proposed the COVID-19 detection model to classify the COVID-19 and healthy patients that claim the higher accuracy. However, such a baseline model suffers from three major limitations. First, these models are highly imbalanced thus, the reported results in baseline studies may not be applicable on a wider scale. Secondly, the researchers used only one modality dataset that was the patient sound modality or cough modality. Third, the overall accuracy of a model is low. Finally, most of the classification models have been employed using traditional ML approaches. This section aims to critically analyze the obtained results and justifies why our proposed MMFFT technique outperformed the single modality dataset and existing baseline techniques. In addition to that this section provides justification that why augmented data could not bring further improvement in the results. Finally, this section also provides the error analyses of misclassified instances by our proposed MMFFT technique.

To overcome the issue of the baseline classification model, we proposed the MMFFT technique to classify COVID-19 and healthy patients using the publicly available COSWARA sound modalities dataset. The experimental results showed the proposed MMFFT technique is effective to classify COVID-19 and healthy patients. Furthermore, the experimental results showed that the proposed multi-modal and feature fusion-based technique is more effective to single modality datasets. In addition, the experimental results showed that our proposed MMFFT technique outperform by achieving 96% accuracy, as compared to four existing baseline techniques. Finally, the COSWARA dataset is highly imbalanced, and thus, we applied the audio augmentation technique to make the dataset balanced and to evaluate whether augmented balanced data improve the classification results Moreover, our experiments showed that the augmentation of data improves the overall performance on classification results.

The finding single modality dataset can classify the COVID-19 and healthy patient 88% to 96% correctly. However, such a single modality dataset can show an error of 4% to 12%. The possible reason for this error may be due to the inability of the features to produce the discriminative and representative pattern for COVID-19 and healthy patients. The experimental result of single modality models were just the right models and these were neither overfitted nor under fitted. However, the results still need further improvement in many modalities. As for COVID-19 symptoms may vary patient to patient [36] and, 45% of COVID-19 patients have breathlessness symptoms, 14% have severe respiratory dysfunction, and 41.7% with voice, swallow, and laryngeal sensitivity. Therefore, we should not rely on a single modality to detect the COVID-19.

Several recent studies have proven that the multi-modality and feature fusion yielded promising results in the field of medicine and most of the researcher studies suggest utilizing the multi-modality techniques to obtain robust results. Therefore, to further improve the classification results obtained from single modality feature fusion-based technique. The MMFFT technique with multi-modality showed promising results compare to single modality results. The number of misclassified instances was reduced significantly. The error was only 4%. The possible reason for improved performance was that we have implied multi-modality with feature fusion-based technique to increase the diversity and fuse the diverse information to the reliabilities, robustness, and improve generalization. Though, the proposed MMFFT technique performed better than a single modality. However, we noticed that the COSWARA dataset was imbalanced. The main cause of non-significantly performances was maybe highly imbalanced class problem [37].

As per our hypothesis, the proposed MMFFT technique performed better than a single modality. Moreover, the COSWARA dataset is imbalanced. In previous studies, it is reported that augmentation techniques can improves classification overall accuracy. Therefore, we used the audio augmentation techniques. The fusion of all nine modalities fed to LSTM with audio augmentation showed improvement of 2% of performances (94% testing accuracy) compared to experimental setting-II (96% training accuracy). Precision (1%), recall (2%) and F1-score by (2%). The possible reason for improved performance was that we have implied multi-modality with feature fusion-based technique to increase the diversity and fuse the diverse information to the reliabilities, robustness, and improve generalization.

We are intent to improve the overall performance of MMFFT by using audio Generative Adversarial Network (GAN) techniques by generating the COVID-19 audios sample to make it balance.

Our experimental results shown that the proposed MMFFT technique outperformed on four baseline techniques [16,18,22,28]. In these baseline studies, the authors have worked on multi-modalities to classify the COVID-19 and healthy patients, still there are some limitations. Firstly, the authors have not deal with feature fusion techniques to make a model generalized. The generalized model is more reliable, stable, and accurate while dealing with feature fusion techniques. Secondly, the authors [18,22] have used Random Forest (RF) classifier and achieved 66.74% and 87.5%. However, The Random Forest (RF) classifier is slower than the LSTM and it can be a slowdown performance of classification when RF has a large number of a tree. In previous studies, the author [16] has used CNN model on Crowdsourced dataset on two modality such as cough and breathe and achieved AUC of 80.46%. Moreover, by using the same dataset, the author [28] applied VGGish algorithms and obtained 80% AUC respectively. In addition to this, the authors have not deal with dataset imbalance [22].

To overcome the limitations, we have employed LSTM algorithm on COSWERA dataset. In this study, we have implied feature fusion on multi-modalities with feature fusion-based technique and improved 17% accuracy as compared to previous studies mention above. Our experimental results shown a high correlation MFV and classification accuracy. In addition, the proposed technique significantly improves the overall performance on multi-modalities with augmentation techniques. Setting II and Setting III Tab. 3 proved that our proposed MMFFT technique can be used as a secondary tool to classify the healthy as well as COVID-19 patients without violating the social distancing rule.

The clinical significance of our proposed studies is that our proposed methodology, MMFFT can be effective in enabling multi-modality-based technology solution for point-of-care detection of COVID-19, resulting into quick detection of the COVID-19. This method provides COVID-19 detection results easily, within 2-3 min, without violating social distance. The proposed model can be integrated into any android app to detect the COVID-19 within a minute. Across the world, anyone could use the COVID-19 app and take great advantage of technology. Moreover, this research will provide new directions to researchers who will pursue research on COVID detection.

The major limitation of this proposed MMFFT is that have balanced the dataset by using data augmentation to avoid the biasness. However, data augmentation is a synthetic data generation which is not good enough in term of contextual. Therefore, this COSWARA dataset was not enough for COVID-19 patient.

6 Conclusion and Future Directions

This study proposed an effective MMFFT technique to classify the healthy and COIVD-19 patients from multi-modality audio files using the COSWARA dataset. In multi-modality, we used nine different modalities and from each modality, five different features were extracted. These features were then fused to create a super MFV. The super MFV was then fed an input to LSTM and algorithms for the classification purpose. Our experimental results showed that our proposed technique outperformed by achieving the accuracy of 96% in both classifiers LSTM and improved the 17%–20% accuracy from the four baseline techniques. Furthermore, the dataset which we used for experiments was highly imbalanced. Thus, we employed audio augmentation techniques to overcome the class imbalance issue. We evaluated our proposed technique on both balanced and imbalanced data and found that our proposed technique showed to improve the overall performance of with augmentation. Our promising results show that the proposed MMFFT technique can be utilized as a secondary tool for classifying the health as well as COVID-19 patients without violating the social distancing rule. Moreover, it can be adopted in many other application areas of audio processing and classification including, sentimental analysis, gender classification, speaker identification, etc. In future, we will design an automatic diagnosis tool for COVID-19 from spectrogram using CycleGAN and Transfer Learning.

Funding Statement: This research has been financially supported by University Malaysia Sabah.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] B. Akarsu, D. Canbay Özdemir, D. Ayhan Baser, H. Aksoy, İ Fidancı *et al.,* "While studies on COVID-19 vaccine is ongoing, the public's thoughts and attitudes to the future COVID-19 vaccine," *Int. Journal of Clinical Practice*, vol. 75, no. 4, pp. e13891, 2021.

[2] S. Denford, K. Morton, J. Horwood, R. de Garang and L. Yardley, "Preventing within household transmission of covid-19: is the provision of accommodation to support self-isolation feasible and acceptable?," *BMC Public Health*, vol. 21, no. 1, pp. 1–13, 2021.

[3] O. P. Mehta, P. Bhandari, A. Raut, S. E. O. Kacimi and N. T. Huy, "Coronavirus disease (COVID-19): comprehensive review of clinical presentation," *Frontiers in Public Health*, vol. 8, pp. 1034, 2021.

[4] T. Jartti, L. Jartti, O. Ruuskanen and M. Söderlund, "New respiratory viral infections," *Current Opinion in Pulmonary Medicine*, vol. 18, no. 3, pp. 271–278, 2012.

[5] M. R. Rahman, M. A. Hossain, M. Mozibullah, F. Al Mujib, A. Afrose *et al.,* "CRISPR is a useful biological tool for detecting nucleic acid of SARS-CoV-2 in human clinical samples," *Biomedicine & Pharmacotherapy*, vol. 140, pp. 111772, 2021.

[6] D. M. Wallace and A. Sweetman, "Comorbid sleep apnea, post-traumatic stress disorder, and insomnia: underlying mechanisms and treatment implications—a commentary on El Solh et al.'s impact of low arousal threshold on treatment of obstructive sleep apnea in patients with post-traumatic stress disorder," *Sleep and Breathing*, vol. 25, no. 2, pp. 605–607, 2021.

[7] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. of Advances in Neural Information Processing Systems*, Lake Tahoe, Nevada, US, pp. 1097–1105, 2012.

[8] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[9] O. Yıldırım, P. Pławiak, R. -S. Tan and U. R. Acharya, "Arrhythmia detection using deep convolutional neural network with long duration ECG signals," *Computers in Biology and Medicine*, vol. 102, pp. 411–420, 2018.

[10] M. Scarpiniti, D. Comminiello, A. Uncini and Y. -C. Lee, "Deep recurrent neural networks for audio classification in construction sites," in *Proc. of 28th European Signal Processing Conf. (EUSIPCO)*, Amsterdam, NL, pp. 810–814, 2021.

[11] N. C. Codella, Q. -B. Nguyen, S. Pankanti, D. A. Gutman, B. Helba *et al.,* "Deep learning ensembles for melanoma recognition in dermoscopy images," *IBM Journal of Research and Development*, vol. 61, no. 5, pp. 1–15, 2017.

[12] L. Brabenec, J. Mekyska, Z. Galaz and I. Rektorova, "Speech disorders in Parkinson's disease: Early diagnostics and effects of medication and brain stimulation," *Journal of Neural Transmission*, vol. 124, no. 3, pp. 303–334, 2017.

[13] J. H. Tan, H. Fujita, S. Sivaprasad, S. V. Bhandary, A. K. Rao *et al.,* "Automated segmentation of exudates, haemorrhages, microaneurysms using single convolutional neural network," *Information Sciences*, vol. 420, pp. 66–76, 2017.

[14] S. Wang, B. Kang, J. Ma, X. Zeng, M. Xiao *et al.,* "A deep learning algorithm using CT images to screen for corona virus disease (COVID-19)," MedRxiv, 2020.

[15] P. Bagad, A. Dalmia, J. Doshi, A. Nagrani, P. Bhamare *et al.,* "Cough against COVID: Evidence of cOVID-19 signature in cough sounds," arXiv preprint arXiv:2009.08790, 2020.

[16] H. Coppock, A. Gaskell, P. Tzirakis, A. Baird, L. Jones *et al.,* "End-2-End COVID-19 detection from breath & cough audio," arXiv preprint arXiv:2102.08359, 2021.

[17] A. Imran, I. Posokhova, H. N. Qureshi, U. Masood, S. Riaz *et al.,* "AI4COVID-19: AI enabled preliminary diagnosis for cOVID-19 from cough samples via an app," arXiv preprint arXiv:2004.01275, 2020.

[18] N. Sharma, P. Krishnan, R. Kumar, S. Ramoji, S. R. Chetupalli *et al.,* "Coswara–a database of breathing, cough, and voice sounds for cOVID-19 diagnosis," arXiv preprint arXiv:2005.*10548*, 2020.

[19] A. Anupam, N. J. Mohan, S. Sahoo and S. Chakraborty, "Preliminary diagnosis of COVID-19 based on cough sounds using machine learning algorithms," in *Proc. of Int. Conf. on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, pp. 1391–1397, 2021.

[20] H. F. Nweke, Y. W. Teh, G. Mujtaba and M. A. Al-Garadi, "Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions," *Information Fusion*, vol. 46, pp. 147–170, 2019.

[21] R. Jahangir, Y. W. Teh, H. F. Nweke, G. Mujtaba, M. A. Al-Garadi *et al.,* "Speaker identification through artificial intelligence techniques: A comprehensive review and research challenges," *Expert Systems with Applications*, vol. 171, pp. 114591, 2021.

[22] D. Grant, I. McLane and J. West, "Multiclass sound event detection for respiratory disease diagnosis," *The Journal of the Acoustical Society of America*, vol. 148, no. 4, pp. 2748–2748, 2020.

[23] G. Murtaza, L. Shuib, A. W. A. Wahab, G. Mujtaba, H. F. Nweke *et al.,* "Deep learning-based breast cancer classification through medical imaging modalities: State of the art and research challenges," *Artificial Intelligence Review*, vol. 53, pp. 1–66, 2019.

[24] E. E. -D. Hemdan, M. A. Shouman and M. E. Karar, "Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images," arXiv preprint arXiv:2003.11055, 2020.

[25] I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: Automatic detection from x-ray images utilizing transfer learning with convolutional neural networks," *Physical and Engineering Sciences in Medicine*, vol. 43, pp. 1, 2020.

[26] J. Shuja, E. Alanazi, W. Alasmary and A. Alashaikh, "Covid-19 open source data sets: A comprehensive survey," *Applied Intelligence*, vol. 51, pp. 1–30, 2020.

[27] M. Wasikowski and X. -w. Chen, "Combating the small sample class imbalance problem using feature selection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1388–1400, 2009.

[28] C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthanasombat *et al.,* "Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data," in *Proc. of the 26th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, Virtual Event, CA, USA, pp. 3474–3484, 2020.

[29] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," arXiv preprint arXiv:1708.04552, 2017.

[30] S. Chen, E. Dobriban and J. H. Lee, "A group-theoretic framework for data augmentation," *Journal of Machine Learning Research*, vol. 21, no. 245, pp. 1–71, 2020.

[31] L. Nanni, G. Maguolo and M. Paci, "Data augmentation approaches for improving animal audio classification," *Ecological Informatics*, vol. 57, pp. 101084, 2020.

[32] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[33] Y. Gong, Y. -A. Chung and J. Glass, "PSLA: Improving audio event classification with pretraining, sampling, labeling, and aggregation," arXiv preprint arXiv:2102.01243, 2021.

[34] S. Hochreiter and J. Schmidhuber, “LSTM can solve hard long time lag problems,” *Advances in Neural Information Processing Systems*, vol. 9, pp. 473–479, 1997.

[35] P. Bojanowski, E. Grave, A. Joulin and T. Mikolov, “Enriching word vectors with subword information,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135–146, 2017.

[36] M. G. Ison and F. G. Hayden, “Viral infections in immunocompromised patients: what’s new with respiratory viruses?” *Current Opinion in Infectious Diseases*, vol. 15, no. 4, pp. 355–367, 2002.

[37] M. Ibrahim, M. Torki and N. El-Makky, “Imbalanced toxic comments classification using data augmentation and deep learning,” in *Proc. of 17th Int. Conf. on Machine Learning and Applications*, Orlando, FL, USA, pp. 875–878, 2018.