COVID-19 detection from cough audio signals using a Recurrent Convolutional Neural Network

Received xxxxxx
Accepted for publication xxxxxx
Published xxxxxx

Abstract

In March 2020, the World Health Organization declared COVID-19 to be a global pandemic. To date it has caused the deaths of over 6 million people worldwide. Gold standard methods for diagnosis of the disease remain challenging for pandemic control due to the need for a physical examination, which cannot be performed remotely. Biomedical research in COVID-19 detection using cough audio signals has gained significance due to the potential of publicly available datasets to provide a vast amount of data for the development and improvement of AI screening tools and deep learning algorithms. This study presents a signal processing system for COVID-19 detection based on participants' cough audio recordings. Our main objective was to investigate the system's ability to identify COVID-19 and make performance comparison to conventional AI-based methods. The system involves cough event detection and segmentation, followed by COVID-19 classification from the detected cough events using a Recurrent Convolutional Neural Network. Audio recordings were collected from 664 participants using publicly available datasets acquired via mobile platforms. We defined exclusion criteria to overcome the limitations of current datasets. The detection of cough events from the audio recordings yielded an UAR and AUC of 92.5% and 98.2% respectively. For COVID-19 classification, this approach achieved an UAR and AUC of 71.7% and 77.4% respectively. The results of the study emphasize the significant possibility of utilizing cough-based signal processing systems for COVID-19 diagnosis. Such a method of detection is rapid, user-friendly, non-intrusive, and can be performed remotely, which is essential during pandemic situations.

Keywords: COVID-19, cough sounds, audio signal processing, CNN, Recurrent CNN

1. Introduction

Viral respiratory diseases are endemic and have a significant impact on morbidity and mortality globally [1]. They are caused by respiratory viruses (RVs) which commonly affect the upper or lower respiratory tract. One such virus is SARS-COV-2, the cause of coronavirus disease 2019 (COVID-19). In March 2020 the World Health Organization (WHO) declared it to be a global pandemic, and is considered responsible for the deaths of over 6 million people [2] to date.

The typical methods for detecting COVID-19 include the reverse transcription polymerase chain reaction (RT-PCR) and the rapid antigen test (RAT), where a sample is taken from the individual's nose or throat to identify the RNA components (in PCR) and nucleoproteins (in RAT) of the virus [3]. Although considered the gold standard, the PCR test remains a challenging method for pandemic control because it requires suspected carriers to breach isolation rules, thus contributing to the spread of the virus. By contrast, RAT can be performed at point-of-care (POC) or even self-administered at home, however, a special diagnosis kit is still needed and the results are less accurate than the PCR test [4]. These limitations underscore the need for alternative methods to detect/diagnose COVID-19, which is crucial for the eradication of the disease and treatment for those who are infected.

One of the most common symptoms of COVID-19 is a dry cough [5]–[7]. Coughing is intricately related to changes in the respiratory system physiology, such that different

pathological conditions can cause the glottis to behave differently, which impacts the acoustics of the cough [8] and can convey information related to the disease. This suggests that screening tool for COVID-19 based on cough sounds and obtained from mobile platforms (i.e., smartphones) could overcome the limitations of current diagnostic methods since it could be done easily, non-invasively and remotely.

Since the outbreak of the pandemic, a number of datasets consisting of cough audio recordings have been made available (Coswara [9], COUGHVID [10], Corona Voice Detect [11], COVID-19 Sounds [12] etc.). Acoustic data have been collected from mobile applications, thus enabling people from all over the world to record and upload their cough sounds. Most datasets also include information on COVID-19 status (positive/negative), age, sex, current symptoms, etc. Thus, extensive research has been initiated to develop AI-based screening tools using cough acoustic data [13]–[17]. Currently, three different challenges have been set up where participants present COVID-19 classification models and report the predictions on blind test-sets. In the Diagnosis of COVID-19 using Acoustics (DICOVA) challenge [18], out of 29 teams, 9 achieved Area Under the ROC Curve (AUC) scores above 80%. First place scored 87% [19]. In the INTERSPEECH 2021 Computational Paralinguistics Challenge (ComParE) [20], 19 teams submitted their test set predictions. The best score was 75.9% Unweighted Average Recall (UAR) [21]. The best score on the second DICOVA challenge was an AUC of 82% out of 21 teams [22]. These results underscore the high potential for COVID-19 diagnosis using AI-based screening platforms.

Although crowdsourcing enables easy collection of large amount of data, there is limited control over the data validity [7]. For example, we found that many participants reported to be COVID-19 positive on questionnaire but their PCR test was dated up to several months prior to the audio recording. All these participants needed to be excluded given the uncertainty of their COVID-19 status. In the COVID-19 Sounds dataset, the organizers addressed this issue by indicating whether the PCR test had been within 14 days prior to the audio recording or not [12]. Some datasets provide the dates of the PCR test and the audio recording, but in others, this information was not even requested in the questionnaire. In most COVID-19 diagnosis studies using crowd-sourced datasets, there is no differentiation between participants based on their PCR test date, and no method to tackle cases in which this information was not provided. Defining exclusion criteria is one way to overcome these data control limitations.

In addition, most cough audio recordings in these datasets also contain non-cough sounds such as speech, breathing, background noise, etc. This means that in classification tasks using cough signals, a cough detection step may be needed to eliminate the non-cough segments of the audio recording before training the classification model. This step usually consists of a simple detection algorithm based on amplitude or energy threshold [7] which is sufficient for distinguishing the cough from the background noise or silent segments in the recording. However, to isolate coughs from higher intensity sounds (e.g., speech and breathing); a more robust algorithm needs to be considered, in conjunction with an evaluation of its contribution to the overall system.

Audio signals for cough analysis are typically analyzed with Convolutional Neural Networks (CNN), especially in the case of COVID-19 detection [23]. The main idea is to divide the audio signal into short time frames, extract the spectrogram from a group of consecutive overlapping frames using a short time Fourier transform (STFT) on each frame to obtain a 2D time-frequency representation of the data. Each 2D image represents a segment of the audio signal and constitutes the input layer to the CNN model that yields decision for that segment. The disadvantage of this approach is that the model can only learn features within each segment and does not take the time variant information between different segments of the entire audio signal into account. To overcome this constraint, Rajan et al. used the entire audio recording as the input layer to their network, where each audio recording was set to a length of 30 sec [24]. However, for crowd-sourced datasets, the audio recordings are dependent on each participant and can vary in size. Some studies have combined CNN with a Recurrent Neural Network (RNN) architecture for time sequential learning of the data. Nevertheless, the decision was made for each segment; in other words, the time sequential learning did not cover the entire audio signal [25], [26]. A different approach is based on Recurrent CNN (RCNN) where all the spectrogram-segments are concatenated along a third axis like consecutive images in a video. The model was able to extract information within each segment along with the time-variant information between segments of the audio signal [27]–[29]. To the best of our knowledge, this method has not been applied to COVID-19 classification.

The goal of the current study was to develop a signal-processing system for the detection of COVID-19 from participants' cough audio recordings extracted from crowd-sourced datasets using mobile platforms. To do so, we applied in-house subject exclusion criteria to improve the validity of the raw datasets. The system is composed of cough event detection and segmentation algorithm for the removal of non-cough sounds from the audio recording, followed by COVID-19 classification using a RCNN model. The RCNN enabled to utilize information from the entire audio signal instead of individual segments. We compared the RCNN model results to a CNN model to evaluate whether sequential learning between segments improved system performance. In addition, we trained both models with and without cough event detection and segmentation, to

assess whether removal of non-cough sounds improved classification results.

2. Methods

2.1. Data collection

Cough audio recordings were acquired from three crowd-sourced datasets: 1) The Coswara dataset of the Indian Institute of Science [9] which contains audio recordings from 2,746 subjects. Each subject recorded two cough audio signals (heavy and shallow), resulting overall in 5,492 recordings. 2) The COUGHVID dataset of the Swiss Federal Institute of Technology in Lausanne [10] which contains cough audio recordings from 27,550 subjects. 3) The Voca dataset of voca.ai and Carnegie Mellon University [11] which contains audio recordings from 2,954 subjects. For all datasets, recordings were acquired via web/mobile platforms in WAV, WEBM or OGG formats at a sampling frequency of 44.1 or 48 kHz based on the subject's device. Each subject was requested to make a recording of several voluntary coughs for a few seconds and to fill out a questionnaire indicating his or her age, gender, COVID-19 status (positive/negative) and symptoms. Each dataset also contains additional information (Table I) such as COVID-19 diagnoses by expert pulmonologists who listened to each recording, an audio quality check and the subjects' PCR test date.

2.2. Subjects' exclusion criteria

The initial database underwent subject exclusion based on the following criteria (Fig. 1): 1) Metadata validity, i.e., any subject with missing or unreliable information (e.g., age < 0) was removed. 2) Ground truth validity. Training a model relies on its ground truth labels and for COVID-19 diagnosis, RT-PCR is the gold standard [1]. Therefore, participants were requested to provide their COVID-19 status based on PCR results. In addition, time elapsed between the PCR test and audio recording also needed to be verified because after a certain amount of time, test results are no longer reliable. Therefore, subjects who did not provide a PCR test date or delayed long enough with the audio recording were excluded. For the COUGHVID dataset, this information was not requested in the questionnaire, so ground truth validity was assessed through a different approach; namely, subjects were only included if there was an agreement between their own self-report and all the pulmonologists' diagnoses. For the Voca dataset, no subject from the negative class provided a PCR test date. Therefore, we excluded subjects who had symptoms at the time of the recording. 3) Audio quality: subjects with low audio quality were removed.

The database specifications after subject exclusion can be seen in Table II. The database consisted of 664 subjects (177 from Coswara, 257 from COUGHVID and 230 from Voca), resulting overall in 832 audio recordings (345 from Coswara).

Table I
DATABASE INFORMATION PER DATASET

	COVID-19 status diagnosis by expert pulmonologists	audio quality check	PCR test date
Coswara		V	V
COUGHVID	V	V	
Voca			V

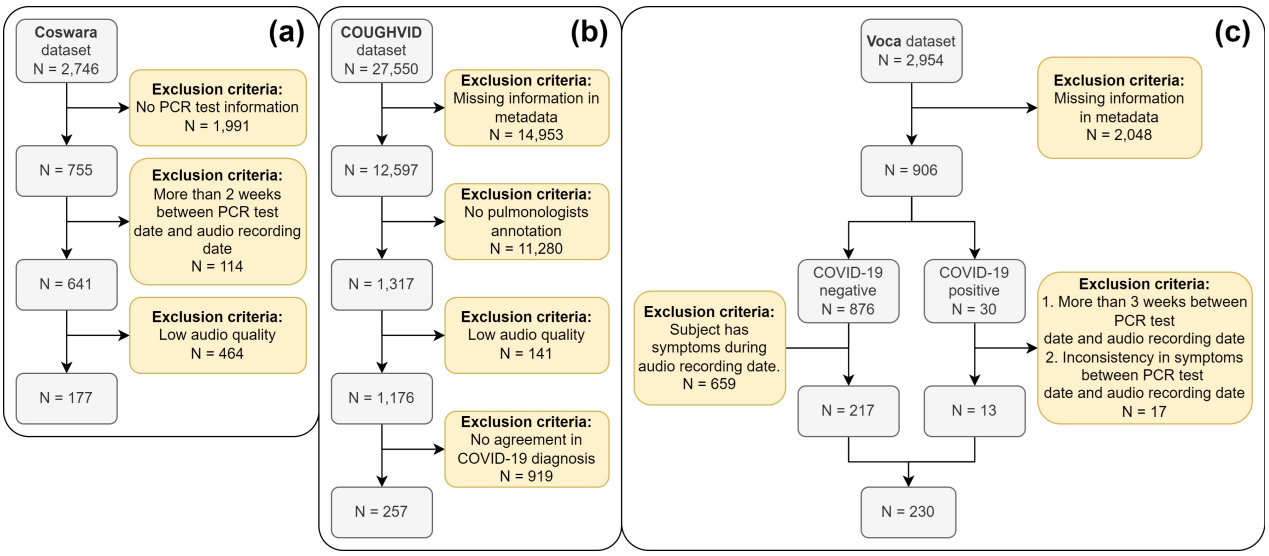

Fig. 1 – Subjects' exclusion criteria for each dataset: (a) Coswara. (b) COUGHVID. (c) Voca.

Table II
Database specification after subject exclusion

	All database	Coswara dataset	COUGHVID dataset	Voca dataset
No. subjects	664	177	257	230
No. Subjects (positive class)	235 (35.4%)	125 (70.6%)	97 (37.7%)	13 (5.7%)
No. subjects (negative class)	429 (64.6%)	52 (29.4%)	160 (62.3%)	217 (94.3%)
No. recordings	832	345	257	230
No. recordings (positive class)	355 (42.7%)	245 (71%)	97 (37.7%)	13 (5.7%)
No. recordings (negative class)	477 (57.3%)	100 (29%)	160 (62.3%)	217 (94.3%)
Age	34.8 ± 14.8	38.2 ± 16.6	32.7 ± 13.1	34.6 ± 14.6
No. male subjects	442 (66.6%)	106 (59.9%)	175 (68.1%)	161 (70%)
No. female subjects	221 (33.3%)	71 (40.1%)	81 (31.5%)	69 (30%)

2.3. Cough-based COVID-19 detection system

A simplified block diagram of the overall system for Cough-based COVID-19 detection is shown in Fig. 2. For each audio recording, the entire signal initially underwent a pre-processing step which included background noise removal followed by a feature extraction step that involved the extraction of the spectrogram from the time-domain audio signal for a time-frequency representation. The spectrogram was segmented along the time axis into N segments, each 388 msec in length. In the cough event detection step, for each segment, a decision was made whether that segment contained a cough event or not. Only segments containing cough events were preserved for further processing, resulting in K segments ($K \leq N$). In the COVID-19 classification step, the RCNN model was used to analyze the time-frequency information within each spectrogram-segment using CNN model, and the time-variant information between segments using RNN model, followed by a final positive/negative COVID-19 decision for each audio recording.

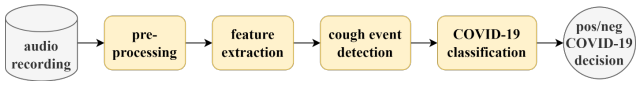

Fig. 2. Simplified block diagram of the cough-based COVID-19 detection system.

2.4. Pre-processing

For each audio recording, the following steps were performed: downsampling from 44.1/48 to 16 kHz, DC removal by subtracting the average value, amplitude normalization by dividing by the maximum absolute value, and attenuation of background noise using magnitude spectral subtraction. The last step was implemented to reduce the influence of the recording environment on the cough sounds and hence on the COVID-19 classification results. The algorithm was based on subtraction of the spectral magnitudes of the signal and the background noise, where the noise was estimated from low-energy part of the signal. For further details, see [30]. An example of the pre-processing step on an audio signal is presented in Fig. 3.

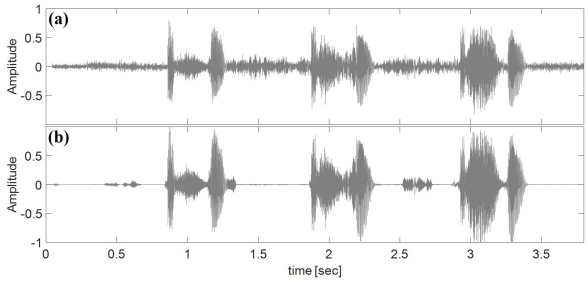

Fig. 3. Pre-processing of an audio recording. (a) Raw audio signal. (b) Audio signal after pre-processing.

2.5. Feature extraction

After pre-processing, the mel-spectrogram was extracted from the audio signal followed by segmentation into N segments along the time axis (Fig. 4). First, each recording was divided into frames of length 128 samples each (8 msec) to maintain stationarity with 64 samples (50%) overlap between frames. Each frame was converted to the frequency domain using a one-sided short time Fourier transform (STFT) with a Hann window and 128 sample zero padding. Then, the frequency axis was converted from a linear to a mel scale using 64 filter banks which is better suited to the way humans perceive frequencies and is widely used in deep learning models. This resulted in a mel-spectrogram of size $64 \times L$ where L represents the length of the time-domain audio recording. The mel-spectrogram was segmented on the time axis into N segments of length 96 samples each (388 msec) and 72 samples (75%) overlap between segments. Overall, this resulted in a $96 \times 64 \times N$ feature matrix for each subject (96 time steps, 64 frequency filter banks and N segments). A segment length of 388 msec was chosen to contain an entire cough event within a single segment while preserving good resolution between segments for the cough detection system. In other words, it produced a separation

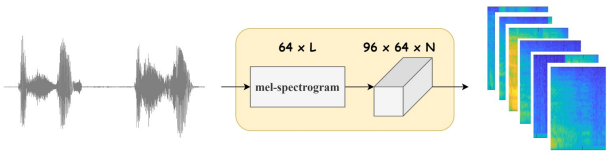

Fig. 4 - Block diagram of the feature extraction method.

into segments containing only cough or only non-cough sounds.

2.6. Cough event detection

We implemented a cough event detection system (Fig. 5) using MobileNet-based architecture named Yet Another Mobile Network (YAMNet).

2.6.1. The YAMNet model

YAMNet is a pre-trained CNN, trained on AudioSet dataset, which contains manually annotated audio events from more than 2 million YouTube videos [31]. It employs the MobileNetV1 depthwise separable convolution architecture which is composed of depthwise and pointwise convolutions [32] and can predict audio events from 521 different classes, such as laughter, barking, sirens, etc. The input layer to the network is a 96×64 mel-spectrogram. All weights and biases were saved from the pre-trained model but not fixed, i.e., they were used for initialization and their values were changed during training. The last fully connected (fc) layer was changed to 2×1, followed by a softmax activation function and a weighted cross-entropy cost function as follows:

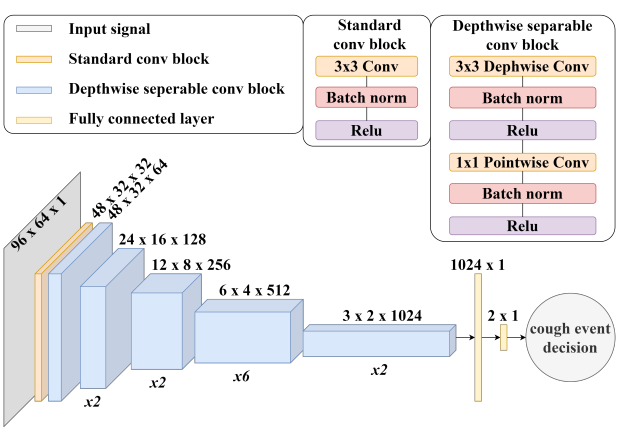

Fig. 5. Block diagram of the cough event detection system.

$$cost = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{2} w_i I_{n,i} \ln\left(y_{n,i}\right) \tag{1}$$

where N is the number of samples (segments) in one mini-batch, w_i is the weight of class i, $I_{n,i}$ is the indicator that sample n belongs to class i, and $y_{n,i}$ is the output from the softmax function for sample n and class i. The weights were calculated as follows:

$$w_i = \frac{N}{2 \cdot \sum_{n=1}^{N} I_{n,i}} \tag{2}$$

where the value of w_i is inversely proportional to the number of samples in class i. This way, the minority class (cough) is penalized more severely for incorrect classification to reduce overfitting towards only correct classification of the majority class (non-cough). Cough/non-cough classification was carried out for each segment. This resulted in a $96 \times 64 \times K$ feature matrix for each subject where K is the number of remaining segments labeled by the model as containing cough events.

2.6.2. Detection smoothing

After the model decision, we added a post-processing step to discard short non-cough intervals that were labelled as coughs. To remove these intervals, we performed 1D morphological opening (composed of erosion and dilation operations) which is mainly used on binary images for the removal of small objects from the image while preserving the shape and size of larger objects [33]. We successfully removed intervals shorter than 150 msec (presumably non-cough) while preserving the length of longer intervals (presumably cough). An example of this cough event detection and smoothing on audio signal is presented in Fig. 6.

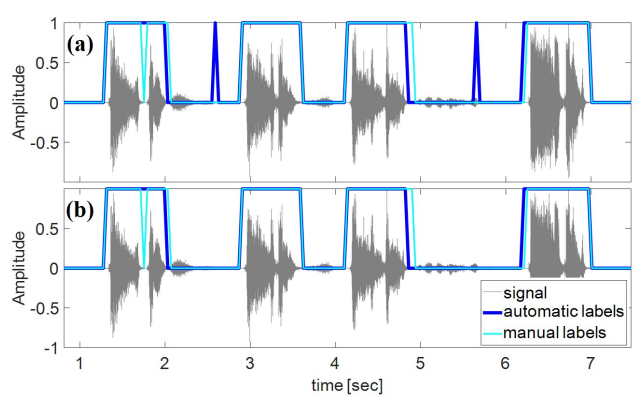

Fig. 6. Audio signal with cough event detection. (a) After the cough event detection. (b) After the detection smoothing.

2.6.3. Model training

The database used for training the model consisted of 171 audio recordings, of which 145 were from the COUGHVID dataset (76 COVID-19 positives and 69 negatives) and 26 from the Voca dataset (13 COVID-19 positives and 13 negatives). Manual labeling of cough events was implemented using Audacity, which is a free open-source digital audio editor software. The overlap between spectrogram segments was set to 90% to increase both the amount of training data and the segmentation resolution. We divided the database (subjects) into 80% train, 10% development (dev) and 10% test. Hyper-parameters tuning was performed using repeated random subsampling validation on the combined train and dev sets. In other words, for each selected hyper-parameters, we combined and split the subjects in the train and dev sets randomly and trained the model. This resulted in the hyper-parameters with the best performance on dev set (Table III). The classification threshold was chosen based on the maximum F1-score on the dev set.

Table III
List of the chosen hyper-parameters for the cough event detection system.

Hyper-parameter	value
Mini-batch size	32
L2 regularization	0.003
Initial learning rate	$3 \cdot 10^{-5}$
Learning rate drop period	After 1 epoch
Learning rate drop factor	0.1

2.7. COVID-19 classification

For the COVID-19 classification, we implemented an RCNN model (Fig. 7). For each subject, output data from the cough event detection system ($96 \times 64 \times K$ feature matrix) was divided into K separate segments, each of size 96×64. The spatial time-frequency information of each spectrogram-segment was embedded using a CNN-VGGish model (Fig. 8) into a 128×1 feature vector. Resulting in K vectors that were the input layers to the RNN-BILSTM model (bidirectional long short-term memory) for analysis of the time-variant information between segments. This model was composed of two LSTM blocks, each with 512 hidden units. One propagates forward in time and the other backwards, where each receives the input vector appropriate to each segment. Hidden layers from each of the two blocks were updated until all segments were passed, and then concatenated,

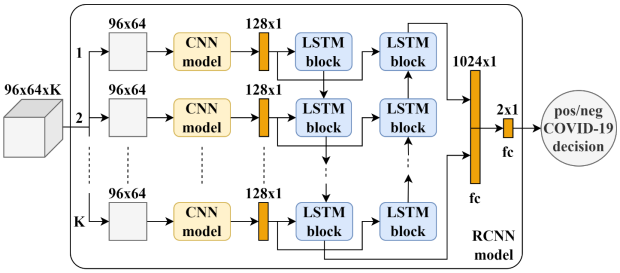

Fig. 7. Block diagram of the RCNN model for COVID-19 classification.

resulting in a 1024×1 fc layer. Dropout was added for regularization which randomly sets the BILSTM weights to zero with 0.1 probability. Then, another 2×1 fc layer was added followed by softmax activation and weighted cross-entropy cost functions (as in the cough event detection). COVID-19 positive/negative classification was carried out for each audio recording.

2.7.1. Adjusting the recording lengths

Each audio recording had a different duration, so that for a feature matrix $96 \times 64 \times K$, K varied across subjects. An LSTM layer is composed of a single block that only updates its parameters as it propagates through the segments. Therefore, K did not need to be fixed for all subjects but only for each mini-batch (where training is done in parallel). During training, for each mini-batch, we zero-padded the feature matrices of all subjects to the largest K. During testing, we set the mini-batch size to be 1 to prevent zero padding and avoid bias caused by the recording lengths.

2.7.2. VGGish model

VGGish (Fig. 8) is a pre-trained CNN from Google [34] designed for audio feature extraction and trained on YouTube-8M [35], a large-scale labeled YouTube video dataset. The input layer to the network was a 96×64 mel-

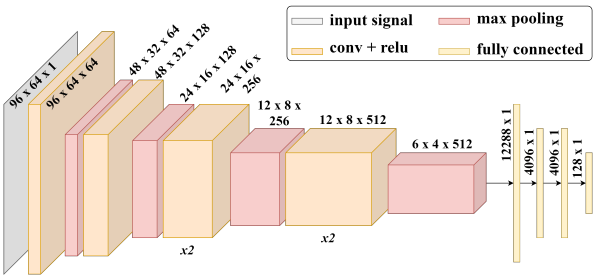

Fig. 8. Block diagram of the VGGish model.

spectrogram. All the weights and biases were saved from the pre-trained model but not fixed (as in the cough event detection).

2.7.3. Model training

We divided the dataset (subjects) into 60% train, 20% dev and 20% test. Hyper-parameters tuning was implemented using repeated random subsampling validation on the train and dev sets (as in the cough event detection), resulting in the hyper-parameters with the best performance on dev set (Table V). The classification threshold was chosen based on maximum F1-score and specificity ≥0.65 on the dev set.

Table V
List of chosen hyper-parameters for the COVID-19 classification system.

Hyper-parameter	Value
Mini-batch size	4
L2 regularization	0.003
Initial learning rate	0.0001
Learning rate drop period	After 1 epoch
Learning rate drop factor	0.7
Number of hidden units per LSTM block	512
Drop out probability	0.1

3. Results

A total of 832 audio recordings were obtained from 664 subjects (235 COVID-19 positives and 429 negatives) in this study for the COVID-19 classification task. Of these, 171 audio recordings were manually labeled for cough events and used for the design and validation of the cough event detection and segmentation system. Statistical significance tests were implemented using corrected resampled t-test proposed by Nadeau and Bengio [36], [37]. In the case of comparison between more than two models, an analysis of variance (ANOVA) was run with same correction method based on Zimmerman et al. [38].

3.1. Cough event detection

Cough/non-cough decision was made for each segment (388 msec). We trained and tested the model over 100 iterations where we shuffled the database (subjects) in each and split into 80% train, 10% dev and 10% test. Performance evaluation is presented for two measurement methods: 1) Per segments (Table IV), where we compared the manually labeled to the predicted segments (system output). 2) Per cough events (Fig. 9), where we measured the overlap (50% and 70%) between the manually labeled and the detected cough event durations. Cough event durations were determined using the timestamps of the onset and offset of each cough event.

Table IV
Cough event detection results per segment on the test set over 100 iterations

Results per segment on test set (**μ±σ** [%])	
Accuracy	93.9 ± 0.4
UAR	92.5 ± 0.7
F1-score	90.5 ± 0.7
Sensitivity	88.6 ± 2
PPV	92.6 ± 1.5
Specificity	96.5 ± 0.8
AUC	98.2 ± 0.1

Evaluation per segments resulted in an accuracy of 93.9%, UAR of 92.5% and AUC of 98.2%. For evaluation per cough events, the F1-score, sensitivity, and positive predictive value (PPV) were calculated. Results for the overlap of 70% yielded an F1-score of 91.1%, a sensitivity of 89.7% and a PPV of 92.6%.

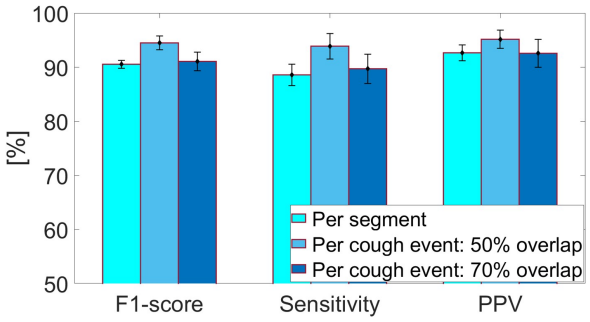

Fig. 9. Cough event detection results per segment and per overlap between the detected vs. the manually labelled cough events on the test set over 100 iterations.

3.2. COVID-19 classification

A COVID-19 positive/negative decision was made for each audio recording. For subjects with two recordings (in the Coswara dataset), the mean value of the positive probability predictions was calculated. We trained and tested the model over 100 iterations where in each we shuffled the database (subjects) and split into 60% train, 20% dev and 20% test. The results on the test set yielded a UAR of 71.7%,

an AUC of 77.4%, a sensitivity of 75.3% and a specificity of 67.1% (Table VI A).

3.2.1. Comparison across different systems

We compared COVID-19 classification results across four different systems (Table VI): A) Our system composed of cough event detection followed by the RCNN classification model. B) The RCNN model without using cough event detection. Instead, we only removed the silent parts from the beginning and end of each recording using a simple energy-based threshold. This resulted in a UAR of 70% and an AUC of 76.6%. C) The CNN-VGGish model (without BILSTM layers). We added a 2×1 fc layer, softmax activation and weighted cross-entropy cost functions to the CNN-VGGish model. Prediction was made for each segment followed by a post-processing step (i.e., taking the mean value between segments) for the final decision for each audio recording. This resulted in a UAR of 73.3% and an AUC of 79.2%. D) The CNN-VGGish model without using cough event detection. This resulted in a UAR of 72.6% and an AUC of 78.8%. All four systems yielded similar results with slightly better performance for the CNN based models, and the models that used cough event detection. Yet, no significant difference was found (p-value > 0.05).

We calculated mean ROC curves for all four systems, as presented in Fig. 10. All four curves are similar with results exceeding random chance. System C (CNN model with cough detection) had slightly higher mean sensitivity values, resulting in overall better AUC scores.

Table VI
COVID-19 classification results. Evaluation per subjects on the test set over 100 iterations. Comparison across different systems: (A) Proposed RCNN model. (B) RCNN model without using cough event detection. (C) CNN-VGGish model. (D) CNN-VGGish model without using cough event detection.

Results on test set: (μ±σ [%])	**(A) RCNN model**	**(B) RCNN model no cough detection**	**(C) CNN model**	**(D) CNN model no cough detection**
Accuracy	70.7 ± 3.4	70.3 ± 3	72.2 ± 3	72 ± 3.1
UAR	71.7 ± 3.7	70 ± 3.3	73.3 ± 3.3	72.6 ± 3.5
F1-score	64.5 ± 4.5	62.2 ± 4.2	66.3 ± 4	65.4 ± 4.2
Sensitivity	75.3 ± 9.5	68.9 ± 8.9	77.2 ± 8.7	74.5 ± 9.3
PPV	56.9 ± 4.1	57.4 ± 4.1	58.6 ± 3.7	58.9 ± 4.1
Specificity	68.1 ± 6.7	71.1 ± 6.2	69.5 ± 6.1	70.7 ± 6.5
AUC	77.4 ± 3.5	76.6 ± 3.1	79.2 ± 3	78.8 ± 2.8

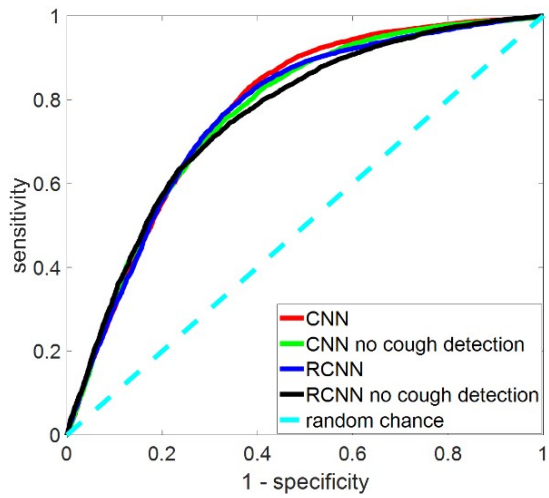

Fig. 10. Mean ROC curve results. Evaluation per subjects on the test set over 100 iterations for the four systems: RCNN/CNN model with/without using cough event detection.

3.2.2. Comparison across subject groups

The COVID-19 classification results from the RCNN model were divided based on age, gender, and symptoms, as shown in Table VII. This was done to analyze the effect of specific sub-groups on system performance. A) Age: the results were compared between subjects above and below the median age of 32 years. The results revealed that the model classified older subjects better although not reaching significance (p-value > 0.05 for both UAR, F1-score and AUC). B) Gender: the model classified male subjects more accurately (p-value < 0.05 for AUC). C) Symptoms: only the Coswara dataset was tested where all subjects stated their COVID-19 status, their PCR test date, and current symptoms at time of the audio recording. The model exhibited high sensitivity (≥80) and low specificity (<30) for both symptomatic and asymptomatic subjects, which is consistent with the results obtained on the Coswara dataset (a sensitivity of 87.9% and a specificity of 24.5%). The accuracy, F1-score and PPV were much higher for subjects with symptoms than for subjects without symptoms.

3.2.3. Positive probability predictions across subject groups

We observed the positive probability predictions on test set obtained from the RCNN model (Fig. 11). Specifically, (A) for class (COVID-19 positive/negative), it produced an overall good separation between classes with higher values

Table VII
COVID-19 classification results. Evaluation per subjects on the test set over 100 iterations. Comparisons per: (A) age, (B) gender, AND (C) symptoms (on Coswara dataset).

Results on test set: (μ±σ [%])	**(A) Age [year]:**		**(B) Gender:**		**(C) Symptoms:**	
	Age < median (~32)	**Age ≥ median (~32)**	**Males**	**Females**	**Symptomatic**	**Asymptomatic**
Accuracy	68.6 ± 5.6	72.6 ± 4.3	73.2 ± 4.3	65.5 ± 5.8	81.2 ± 9.7	57.3 ± 9.4
UAR	70.1 ± 5.8	73.2 ± 4.5	73.6 ± 4.7	67.2 ± 5.6	52.5 ± 15	56.2 ± 8.6
F1-score	61.2 ± 7.1	67.1 ± 6.2	64.6 ± 5.7	63.8 ± 7.1	89.1 ± 6.3	66.6 ± 10.6
Sensitivity	74.4 ± 11.2	76.2 ± 10.5	74.7 ± 10.8	76.3 ± 11.1	89.5 ± 9.1	85 ± 15.6
PPV	53 ± 7.9	60.8 ± 6.6	57.7 ± 5.9	55.7 ± 7.6	89.3 ± 6.3	55.5 ± 9.8
Specificity	65.9 ± 8.5	70.3 ± 7.6	72.4 ± 7.2	58.2 ± 9.9	15.6 ± 28.2	27.3 ± 14
AUC	75.5 ± 5.9	79.3 ± 4.4	80.6 ± 4.3	70.1 ± 7.2	47.1 ± 24.4	58.2 ± 11.3

for the positive class. (B) In terms of age, there was no significant difference between results. (C) For gender, there were lower values for male subjects. (D) For symptoms obtained from the Coswara dataset and divided into four sub-groups (positive/negative with/without symptoms), all four sub-groups were skewed towards higher values. There was slight difference between positive and negative sub-groups with higher values for the positive class in comparison to the negative asymptomatic sub-group. When comparing within the negative class sub-groups, symptomatic subjects had slightly higher values than asymptomatic subjects.

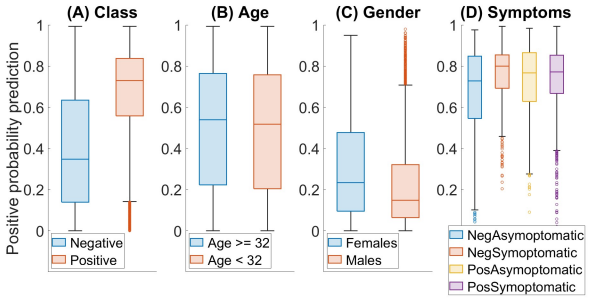

Fig. 11. Positive probability prediction of RCNN model. Evaluation per subjects on test set over 100 iterations for: (A) COVID-19 status; (B) age; (C) gender; (D) symptoms (on the Coswara dataset).

4. Discussion

In this study, a COVID-19 detection system based on cough audio signals was presented. Cough sounds were acquired from three different publicly available crowd-sourced datasets (Coswara, COUGHVID and Voca). We analyzed the audio recordings of 664 subjects after applying our exclusion protocol. Prior training the COVID-19 classification model, we first conducted pre-processing steps involving attenuation of background noise using spectral subtraction. This followed the detection of cough events implemented by CNN-MobileNet model to remove non-cough segments from the audio recordings. The COVID-19 detection was performed using the RCNN based model to be able to utilize information from the entire audio signal during training, for model decision for each recording.

Although currently available crowd-sourced datasets have enabled the rapid rise of research in this area there are still pitfalls that need to be considered in terms of data control. Training a model relies on its ground truth labels and for COVID-19 diagnosis, the PCR test is the gold standard. The participants were requested to provide their COVID-19 status (positive/negative) based on their PCR test results, but this only amounted to self-reports. Without any certification, the subjects' trustworthiness had to be assumed. To overcome this limitation in the future when using crowd-sourced datasets, we strongly suggest requesting the subjects to upload their PCR test certificate.

We removed participants who did not meet our exclusion criteria. We mainly focused on audio quality, relevant information in the metadata, and the time elapsed between the PCR test date and audio recording date. The last criterion is crucial since after a certain amount of time, the test results are no longer reliable because the subject's COVID-19 status may have changed. Subjects who do not provide a PCR test date or waited so long to participate that their test result is no longer reliable should not be immediately included in the database. One disadvantage of this approach is obviously the

curtailment of the size of the initial database. Nonetheless, this step is essential to increase the reliability of the database and define a baseline.

The pre-processing steps including background noise removal must be carried out before further processing. In crowd-sourced datasets, the subject decides where to make the recording. Thus, background noise can differ as a function of the location and lead to bias in the results. For example, people with a medical condition such as COVID-19 may remain stay indoors (e.g. hospital, home, etc.) during the recording as opposed to non-symptomatic individuals who are more likely to be outdoors [39]. We attenuated the background noise from the recordings to provide a good baseline for all audio signals and prevent bias in classification.

A cough event detection step is also necessary if the objective of the model is to learn COVID-19 biomarkers based solely on cough acoustic data. In this case, any other non-cough segments in the recording could turn the algorithm away from its main objective. We evaluated the performance of our cough event detection system segment-by-segment and in terms of overlap between detected and manually labelled cough events. Overall, good results were obtained for both measurement methods with an F1-score and a PPV above 90%. This points to the high capabilities of the model to differentiate between cough and non-cough events. PPV was higher than sensitivity, indicating that the model was better at the identification and removal of true non-coughs than the identification of true coughs.

The evaluation of COVID-19 classification showed that cough event detection improved the results slightly, but the difference was not significant. Thus, suggesting that non-cough segments did not provide additional information for the COVID-19 classification. This could also imply that the model successfully learned cough-based biomarkers and was able to disregard non-cough segments during training. The training time of the COVID-19 classification model was shortened significantly after using cough detection and dropped from 21 to 17 minutes in the RCNN model and from 11 to 8 minutes in the CNN model (on average). Given that non-cough segments did not provide additional information, their removal before training may thus contribute to a faster training process. Future research should expand the cough event detection system to additional tasks such as the detection of breathing and speech events. It can also be used for subject exclusion based on the occurrence of the event in the recording. Subjects for whom the model does not recognize any desired event in the recording should be excluded from the database.

The COVID-19 classification results of our system demonstrated its ability to differentiate between positive and negative labelled subjects. The comparison between the RCNN and CNN models showed that the CNN model was slightly better although the finding did not reach significant difference. The main difference between the two models is that the RCNN learn from the entire audio recording while the CNN made prediction based on individual 388 msec segments. The fact that no significant difference was attained indicate that the model does not need to learn from the entire audio signal and segment-level features are sufficient for obtaining similar results. The CNN model also enables extending the amount of training data (from No. recordings to No. segments times No. recordings), and as observed, reducing training time (by ~10 min). Therefore, considering the results along with final data size, model complexity and running times, training the model based on individual segments is the better choice over training based on the entire recording observation.

The cross- group comparisons indicated no significant impact of subjects' age on the results. On the other hand, the results differed in terms of gender, with more accurate classification for male subjects. This may be explained by the fact that the database contained more male (66.6%) than female subjects (33.3%). As a result, the training model was able to better learn and predict with respect to male subjects. A good approach to reduce this bias would be to balance the number of male/female subjects during training process. A comparison based on symptoms revealed similar positive probability prediction values for subjects who tested positive for COVID-19, whether they had symptoms or not. These values were slightly higher than those in the negative asymptomatic group, but lower than those in the negative symptomatic group. This points to good detection of COVID-19 positives regardless of whether they have symptoms, along with the ability to differentiate them from negative asymptomatic subjects. However, the difference with negative symptomatic subjects was less pronounced.

5. Conclusion

Remote diagnosis of a disease can helps prevent its spread in the early stages and provide better care for those infected. For viral respiratory diseases such as COVID-19, gold standard methods remain challenging given the need for hands-on analysis of the patient, thus promoting further spread of the virus.

This study presented an audio-processing system for the detection of COVID-19 from cough recordings using mobile platforms. The classification results indicate that the algorithm can distinguish between COVID-19 positive and negative individuals. We examined whether our RCNN model (compared to a CNN model) and using cough event detection contributed to performance. We found that the CNN model provided best results with the shortest run time. This indicates that for COVID-19 detection from cough audio signals, training the model using segment-level

features is preferable to training using the entire recording (which impacts the model's complexity and training time). The use of cough event detection proved advantageous due to the removal of non-cough segments that appeared to make no contribution to the results and cut down on training time. A subject exclusion protocol was implemented to overcome data control limitations in current publicly available datasets for COVID-19 diagnosis.

Overall, the findings highlight the potential of cough-based signal-processing systems for COVID-19 detection. This type of diagnosis is fast, easy to use, non-invasive and can be done remotely, which is crucial for monitoring respiratory diseases during a pandemic.

6. Acknowledgment

The authors would like to thank the Data Science Center, Ben-Gurion University of the Negev for its partial support for this study.

7. References

[1] K. Schaffer, A. M. La Rosa, and E. Whimbey, "Respiratory viruses," Infectious Diseases, pp. 1598–1608, 2010, doi: 10.1016/B978-0-323-04579-7.00162-3.

[2] "WHO Coronavirus (COVID-19) Dashboard." https://covid19.who.int

[3] N. Health, "Difference between RT-PCR test and rapid antigen test," Narayana Health Care, Feb. 08, 2022. https://www.narayanahealth.org/blog/difference-between-rt-pcr-test-and-rapid-antigen-test/

[4] R. L. Smith et al., "Longitudinal Assessment of Diagnostic Test Performance Over the Course of Acute SARS-CoV-2 Infection," The Journal of Infectious Diseases, vol. 224, no. 6, pp. 976–982, Sep. 2021, doi: 10.1093/infdis/jiab337.

[5] D. Wang et al., "Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus–Infected Pneumonia in Wuhan, China," JAMA, vol. 323, no. 11, p. 1061, Mar. 2020, doi: 10.1001/jama.2020.1585.

[6] C. Menni et al., "Real-time tracking of self-reported symptoms to predict potential COVID-19," Nat Med, vol. 26, no. 7, Art. no. 7, Jul. 2020, doi: 10.1038/s41591-020-0916-2.

[7] A. Serrurier, C. Neuschaefer-Rube, and R. Röhrig, "Past and Trends in Cough Sound Acquisition, Automatic Detection and Automatic Classification: A Comparative Review," Sensors, vol. 22, p. 2896, Apr. 2022, doi: 10.3390/s22082896.

[8] J. Korpás, J. Sadlonová, and M. Vrabec, "Analysis of the cough sound: an overview," Pulm Pharmacol, vol. 9, no. 5–6, pp. 261–268, Dec. 1996, doi: 10.1006/pulp.1996.0034.

[9] N. Sharma et al., "Coswara -- A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis," Interspeech 2020, pp. 4811–4815, Oct. 2020, doi: 10.21437/Interspeech.2020-2768.

[10] L. Orlandic, T. Teijeiro, and D. Atienza, "The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms," Sci Data, vol. 8, no. 1, Art. no. 1, Jun. 2021, doi: 10.1038/s41597-021-00937-4.

[11] P. Mouawad, T. Dubnov, and S. Dubnov, "Robust Detection of COVID-19 in Cough Sounds," SN COMPUT. SCI., vol. 2, no. 1, p. 34, Jan. 2021, doi: 10.1007/s42979-020-00422-6.

[12] T. Xia et al., "COVID-19 Sounds: A Large-Scale Audio Dataset for Digital Respiratory Screening," presented at the Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), Nov. 2021. Accessed: Nov. 02, 2022. [Online]. Available: https://openreview.net/forum?id=9KArJb4r5ZQ

[13] T. Hoang, L. Pham, D. Ngo, and H. D. Nguyen, "A Cough-based deep learning framework for detecting COVID-19," Annu Int Conf IEEE Eng Med Biol Soc, vol. 2022, pp. 3422–3425, Jul. 2022, doi: 10.1109/EMBC48229.2022.9871179.

[14] X. Jing et al., A Temporal-oriented Broadcast ResNet for COVID-19 Detection. 2022.

[15] H. Coppock, A. Gaskell, P. Tzirakis, A. Baird, L. Jones, and B. Schuller, "End-to-end convolutional neural network enables COVID-19 detection from breath and cough audio: a pilot study," BMJ Innov, vol. 7, no. 2, pp. 356–362, Apr. 2021, doi: 10.1136/bmjinnov-2021-000668.

[16] B. L. Y. Agbley et al., "Wavelet-Based Cough Signal Decomposition for Multimodal Classification," in 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), Dec. 2020, pp. 5–9. doi: 10.1109/ICCWAMTIP51612.2020.9317337.

[17] M. Aly, K. H. Rahouma, and S. M. Ramzy, "Pay attention to the speech: COVID-19 diagnosis using machine learning and crowdsourced respiratory and speech recordings," Alexandria Engineering Journal, vol. 61, no. 5, pp. 3487–3500, May 2022, doi: 10.1016/j.aej.2021.08.070.

[18] A. Muguli et al., "DiCOVA Challenge: Dataset, task, and baseline system for COVID-19 diagnosis using acoustics," Mar. 2021. [Online]. Available: https://ui.adsabs.harvard.edu/abs/2021arXiv210309148M

[19] "DICOVA | IS2021." https://dicova2021.github.io/#results

[20] B. Schuller et al., "The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates," undefined, 2021, doi: 10.21437/interspeech.2021-19.

[21] H. Coppock et al., "A Summary of the ComParE COVID-19 Challenges." arXiv, Feb. 17, 2022. [Online]. Available: http://arxiv.org/abs/2202.08981

[22] N. K. Sharma, S. R. Chetupalli, D. Bhattacharya, D. Dutta, P. Mote, and S. Ganapathy, "The Second Dicova Challenge: Dataset and Performance Analysis for Diagnosis of Covid-19 Using Acoustics," 47th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2022, pp. 556–560, 2022.

[23] A. Ijaz et al., "Towards using cough for respiratory disease diagnosis by leveraging Artificial Intelligence: A survey," Informatics in Medicine Unlocked, vol. 29, p. 100832, Jan. 2022, doi: 10.1016/j.imu.2021.100832.

[24] V. Rajan, A. Brutti, and A. Cavallaro, "ConflictNET: End-to-End Learning for Speech-Based Conflict Intensity Estimation," IEEE Signal Processing Letters, vol. 26, no. 11, pp. 1668–1672, Nov. 2019, doi: 10.1109/LSP.2019.2944004.

[25] M. Chen, X. He, J. Yang, and H. Zhang, "3-D Convolutional Recurrent Neural Networks With Attention Model for Speech Emotion Recognition," IEEE Signal Processing Letters, vol. 25, no. 10, pp. 1440–1444, Oct. 2018, doi: 10.1109/LSP.2018.2860246.

[26] T. Yan, H. Meng, E. Parada-Cabaleiro, S. Liu, M. Song, and B. W. Schuller, "Coughing-Based Recognition of Covid-19 with Spatial Attentive ConvLSTM Recurrent Neural Networks," in Interspeech 2021, Aug. 2021, pp. 4154–4158. doi: 10.21437/Interspeech.2021-630.

[27] C. Li, "Robotic Emotion Recognition Using Two-Level Features Fusion in Audio Signals of Speech," IEEE Sensors Journal, vol. 22, no. 18, pp. 17447–17454, Sep. 2022, doi: 10.1109/JSEN.2021.3065012.

[28] L. Shi, K. Du, C. Zhang, H. Ma, and W. Yan, "Lung Sound Recognition Algorithm Based on VGGish-BiGRU," IEEE Access, vol. 7, pp. 139438–139449, 2019, doi: 10.1109/ACCESS.2019.2943492.

[29] G. Gupta, M. Kshirsagar, M. Zhong, S. Gholami, and J. L. Ferres, "Comparing recurrent convolutional neural networks for large scale bird species classification," Sci Rep, vol. 11, no. 1, Art. no. 1, Aug. 2021, doi: 10.1038/s41598-021-96446-w.

[30] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 27, no. 2, pp. 113–120, Apr. 1979, doi: 10.1109/TASSP.1979.1163209.

[31] J. F. Gemmeke et al., "Audio Set: An ontology and human-labeled dataset for audio events," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Mar. 2017, pp. 776–780. doi: 10.1109/ICASSP.2017.7952261.

[32] A. G. Howard et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications." arXiv, Apr. 16, 2017. doi: 10.48550/arXiv.1704.04861.

[33] J. Serra, "Introduction to mathematical morphology," Computer Vision, Graphics, and Image Processing, vol. 35, no. 3, pp. 283–305, Sep. 1986, doi: 10.1016/0734-189X(86)90002-2.

[34] E. Koh and S. Dubnov, "Comparison and Analysis of Deep Audio Embeddings for Music Emotion Recognition." arXiv, Apr. 13, 2021. doi: 10.48550/arXiv.2104.06517.

[35] S. Abu-El-Haija et al., "YouTube-8M: A Large-Scale Video Classification Benchmark." arXiv, Sep. 27, 2016. doi: 10.48550/arXiv.1609.08675.

[36] C. Nadeau and Y. Bengio, "Inference for the Generalization Error," Machine Learning, vol. 52, no. 3, pp. 239–281, Sep. 2003, doi: 10.1023/A:1024068626366.

[37] I. H. Witten, E. Frank, and M. A. Hall, Data mining: practical machine learning tools and techniques, 3rd ed. Burlington, MA: Morgan Kaufmann, 2011.

[38] D. W. Zimmerman and B. D. Zumbo, "Correction for Nonindependence of Sample Observations in ANOVA F Tests," The Journal of Experimental Education, vol. 60, no. 4, pp. 367–381, 1992.

[39] H. Coppock, L. Jones, I. Kiskin, and B. Schuller, "COVID-19 detection from audio: seven grains of salt," The Lancet Digital Health, vol. 3, no. 9, pp. e537–e538, Sep. 2021, doi: 10.1016/S2589-7500(21)00141-2.