Int. j. inf. tecnol.
https://doi.org/10.1007/s41870-025-02433-z

ORIGINAL RESEARCH

Cross-dataset COVID-19 transfer learning with data augmentation

Bagus Tris Atmaja[2] · Zanjabila[1] · Suyanto[1] · Wiratno Argo Asmoro[1] · Akira Sasou[2]

Received: 29 August 2024 / Accepted: 20 January 2025

Abstract This paper presents a novel cross-dataset transfer learning approach for cough-based COVID-19 detection, enhancing model performance through data augmentation. Our methodology significantly improves results compared to baseline methods. An ablation study highlights the importance of alpha mixup among various hyperparameters in optimizing performance. The final model achieves an unweighted accuracy of 88.19%. Additionally, we provide a comparative summary with previous studies on the same evaluation set to offer insights into cough-based detection methods.

Keywords Cough detection · Cough segmentation · Transfer learning · Data augmentation · COVID-19

1 Introduction

The coronavirus disease that spread at the end of 2019 in China (COVID-19) and in early 2020 over the world is showing the unpreparedness of humans for the pandemic. Although a similar case has occurred previously (Severe acute respiratory syndrome [SARS], Middle East Respiratory Syndrome [MERS]), the response to diagnose the virus in a short time to prevent its spread is not optimal. The gold standard polymerase chain reaction (PCR) test takes time in days and hours. The need for preliminary screening by using other tools is crucial to avoid the spread of the virus. As stated in [1], "The world has been altered with the COVID-19 virus", and living with the virus is a new normal but needs to be prepared since it causes a significant impact on the economy, health, and society [2].

The human voice has the potential as the new blood [3]. For many diseases, a blood test is the main tool to assess the severity. A blood test is also useful for the evaluation of the general condition of the human body. By performing a blood test, the infection of an organ could be checked; the function of certain organs could be monitored. Modern medical imaging techniques (fMRI, CT scan, X-ray) help medical doctors make accurate diagnoses. On the other hand, the use of the human voice for diagnosing particular diseases is limited. Nevertheless, acoustic signals from the human body represent the state of the parts that produce these signals. The sound of the human voice has a potential as the new blood that could be used for diagnosing particular diseases like in blood and imaging.

It has been evidenced that the human voice could be used as the main tool for diagnosing diseases related to voice production: pathological voice detection [4], pertussis [5], asthma [6], and respiratory diseases [7]. Moreover, the use of acoustic analysis has been proven effective for Alzheimer's disease classification (accuracy of 93.30%) [8] and lung disease (accuracy of 98.92%) [9]. These examples show the potency of the human voice for diagnosing diseases, particularly voice-related diseases.

✉ Bagus Tris Atmaja
b-atmaja@aist.go.jp

Zanjabila
zanjabilaabil@gmail.com

Suyanto
suyanto@ep.its.ac.id

Wiratno Argo Asmoro
wiratno@ep.its.ac.id

Akira Sasou
a-sasou@aist.go.jp

[1] Sepuluh Nopember Institute of Technology, Surabaya, Indonesia

[2] National Institute of Advanced Industrial Science and Technology, Tokyo, Japan

Published online: 13 February 2025

Attempts to explore acoustic analysis beyond voice-related diseases have been conducted for COVID-19 detection. Several modalities have been tried, including cough, breathing sounds, and speech or voice (including vowels). The first two modalities contain indicators for the symptoms of COVID-19: continuous cough and shortness of breath. We chose the first modality due to the large available cough data. The cough sounds also showed the highest specificity among other modalities in the previous study [10].

Nowadays, computers are part of daily human equipment, ranging from personal computers to smartphones, which could be used for such acoustic analysis. Given the benefit of diagnosing particular diseases (e.g., COVID-19, dementia, depression), there is a development of an application on a smartphone to detect flare-ups of pulmonary disease [11]. In the future, we predict that smartphones could be furnished with acoustic analysis applications to detect diseases like COVID-19 (Fig. 1). On that day, we could simply ask our smartphone if we have COVID-19 or not for preliminary screening, as in Fig. 1. This application will likely be available not only for COVID-19 but also for other diseases and disorders [11]. In the current stage, applications for classifying the crying of a baby, counting coughs, and detecting anxiety and depression are available in the market. All of these applications are based on acoustic analysis.

To contribute to research on cough-based COVID-19 detection, we aim to investigate three important processing blocks for improving the generalization capability of COVID-19 detection via transfer learning and cross-dataset evaluation. These blocks are cough detection, cough segmentation, and data augmentation. *We would like to know the contribution of these blocks relative to each other compared to the baseline methods without these blocks*. In addition to the main contribution above, we also performed an ablation study to optimize the hyperparameters in the transfer learning stage and a summary of this study with previous studies on the same evaluation (test) dataset.

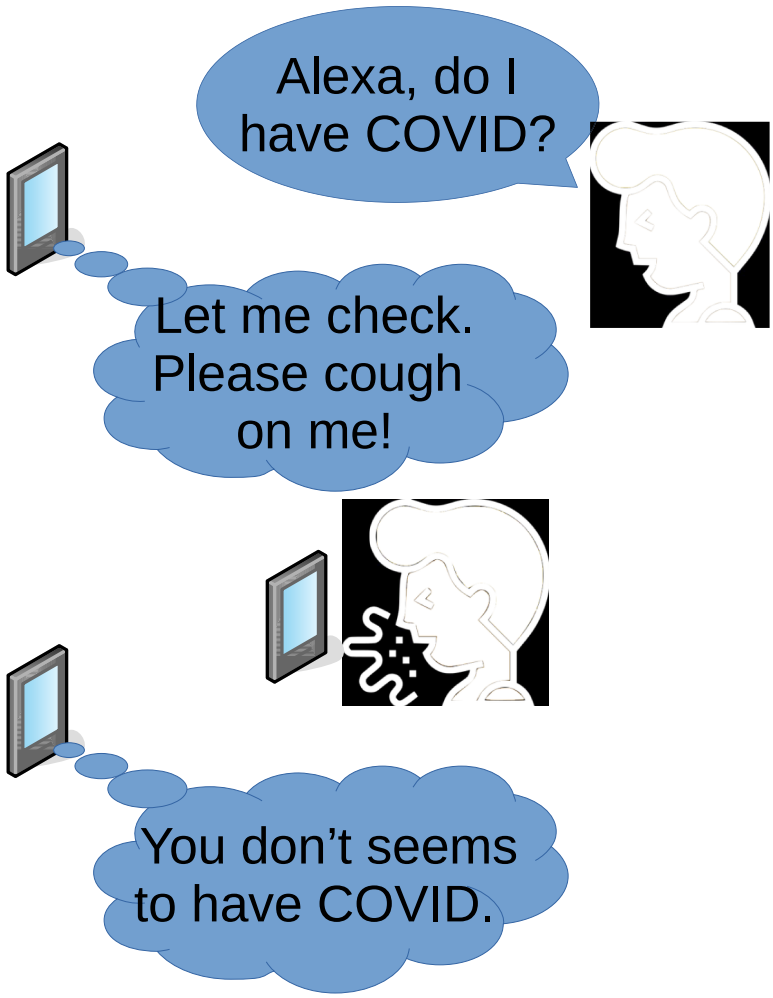

Fig. 1 Illustration of the smartphone application for COVID-19 detection

2 Related work

Research on voice-based COVID-19 detection has been conducted since the spread of the coronavirus disease. Many researchers have proposed different approaches to classify the given data to detect existence of the virus, e.g., using protein sequence data [12] or clinical text data [13]. We focus on classifying human voices into healthy and unhealthy classes for COVID-19 detection, including the use of cough sounds. These approaches could be classified into acoustic features related to cough/COVID-19, models, ensemble methods, transfer learning, and multimodal analysis.

The search of voice biomarkers related to disease has been attempted, for instance the correlation between clinical data parameter and lung sounds [14]. In [15], the authors proposed a framework for COVID-19 biomarkers based on the coordination of speech-production subsystems. Motivated by a unique feature of COVID-19 involving lower and upper respiratory tract inflammation, the authors measure Cohen's effect size between pre- and post-COVID-19 via coordination of respiration (via waveform amplitude) and laryngeal motion (via fundamental frequency and cepstral peak prominence), and coordination of laryngeal and articulatory (via the center of formants). The results show reduced complexity (measured in effect size) between pre- and post-COVID-19 on interview-like voices. This result indicates a possible biomarker for COVID-19 detection via respiratory function captured by acoustic signals. Other studies [16–18] evaluated the existing acoustic features for specific models. For instance, the authors of [17] measured the F-score of 25 acoustic features and found that maximum phonation time (MPT) is the most important feature in their study with /a/ vowel sounds.

The models suitable for COVID-19 prediction also have been proposed, e.g., Hamidi et al. use Hidden Markov Model (HMM) based automatic speech recognition system to analyze the cough signal and determine whether the signal belongs to a sick or healthy speaker [19]. Their proposed method is able to classify dry cough with sensitivity from 85.86% to 91.57%, differentiate the dry cough, and cough COVID-19 symptom with specificity from 5 to 10/%. Hasan et al. [20] applying data analytics and knowledge management approach using Auto- Regressive Integrated Moving Average (ARIMA) model from time-series data to predict the logistics requirements, active cases, positive patients, and death rate of COVID-19.

The use of ensemble methods for COVID-19 detection has been proposed in [21–23]. In [21], the authors proposed an ensemble of convolutional neural network (CNN) classifiers from different acoustic features. In [22], the authors proposed an ensemble-based multi-criteria decision-making (MCDM). These criteria include accuracy, AUC, precision, recall, F1-score, sensitivity, and specificity; the MCDM considers several criteria instead of one. In [23], the authors evaluated ensemble methods using different seed numbers. All the reported results show that ensemble methods can improve the performance of COVID-19 detection.

Transfer learning is now gaining popularity in many fields and has been experimented with audio classification for COVID-19 detection. In [23], the authors transfer the knowledge of the pre-trained model from the AudioSet dataset to the ComParE-CCS dataset. In [10], the authors transfer the knowledge learned from a large audio dataset to the Cambridge COVID-19 sound dataset. Both studies revealed the effectiveness of transfer learning to improve the representation of cough sounds in COVID-19 detection.

The use of multimodal (i.e., cough, vowel, and breathing sounds) in COVID-19 detection has been evaluated in [10, 24]. In [24], the authors extracted local binary patterns and Haralick's features from the spectrogram to analyze the audio textural behavior of cough, breath, and speech sounds. The authors achieved accuracy rates of 98.9% for 2-class and 72.2% for 5-class classification. The authors of [10] reported that a multimodal approach outperformed any single modality approach; breathing achieved the best sensitivity, and cough sounds achieved the best specificity in unimodal evaluations.

Instead of choosing those approaches, we choose to evaluate different pre-processing blocks as continuation of the previous methods [25, 26]. In automatic speech recognition and speech emotion recognition, a little step in pre-processing is important, e.g., extracting silence region [27, 28]. In this study, we evaluated two techniques related to cough and a technique related to general machine learning. The techniques related to cough are cough detection [25, 29] and cough segmentation [26]. Detecting cough will filter out non-cough signals and cough signals with low probabilities. Segmenting cough will split several coughs in a waveform into individual coughs. We argued that this splitting method is more useful than fixed-time splitting (e.g., in [21]) since the region of cough is extracted based on acoustics. The last proposed block is data augmentation to increase the number of samples after cough detection and cough segmentation. The combination of these three blocks was not found in previous studies, as shown in Tables 1, 6.

3 Methods

This method used in this study is based on the previous work [23]. That work only obtained a slight improvement from the original baseline dataset [31], although the authors have employed advanced techniques by utilizing transfer learning, data augmentation, and ensemble methods. We proposed to evaluate three processing blocks to improve the previous results: cough detection, cough segmentation, and data

Table 1 Summary of literature review on voice/cough-based COVID-19 detection (CD: cough detection; CS: cough segmentation; DA: data augmentation)

Ref	Topic	CD	CS	DA
[15]	Framework for biomarkers based on coordination of speech-production subsystems	–	–	–
[16]	Study of using cough sounds and deep neural networks	–	–	–
[17]	Artificial intelligence-based using acoustic parameters	–	–	–
[18]	Acoustic correlates of infection in sustained vowels	–	–	–
[19]	Hidden Markov Model-based automatic speech recognition system	✓	–	–
[21]	Ensemble of CNN classifiers from different acoustic features	–	✓	–
[23]	Transfer learning (TL) and data augmentation	–	–	✓
[30]	Deep learning with large aggregated datasets	✓	–	–
[26]	Cough detection and segmentation methods	✓	✓	–
This study	Cross-dataset TL with cough detection, cough segmentation, data augmentation	✓	✓	✓

augmentation. The method is shown in Fig. 2. Nevertheless, we describe the dataset, classifier, and evaluation metric for a complete understanding of the method.

3.1 Datasets

Three datasets are merged to enable cross-dataset evaluations. The test set is taken from the Computational Paralinguistic Challenge COVID-19 Cough Sub-challenge (ComParE-CCS) 2021 test set to compare our method with previous studies. The rest of the ComParE-CCS data are merged with Coswara and COUGHVID datasets. The description of each dataset is shown below.

Coswara

Coswara is a database of breathing, cough, and voice sounds for COVID-19 diagnosis [32]. Among the three types of sound (speech, cough, and breathing), we only used the cough sounds. These sounds are recorded through worldwide crowdsourcing using a website application. Although the recording is targeted at worldwide participants, about 80% of the speakers came from India. We resampled the original audio files from 48 kHz to 16 kHz. The cough sounds are then manually normalized in a range [−1, 1] and saved again in a WAV format (Fig. 2). The normalization process was conducted by using the Librosa [33] toolkit. The Coswara dataset is retrieved from their GitHub repository with commit ID “401b516".

COUGHVID

COUGHVID is a crowdsourcing dataset for the study of large-scale cough analysis algorithms [29]. The dataset focuses on three features: detection of cough, expert labeling of cough, and high correlation between symptomatic and COVID-19 labels and location of the speakers with high infection rates. We employed both the cough detection model and COUGHVID data in this cough-based COVID-19 detection study. The original sampling rate of 48 kHz was resampled to 16 kHz. The format is also converted from WEBM to WAV with normalization in a range of [−1, 1]. The COUGHVID dataset was retrieved from their Zenodo repository with Version 3.0.

ComParE-CCS

ComParE-CCS is a dataset for cough-based COVID-19 detection [31] originally for the INTERSPEECH 2021 Computational Paralinguistic Challenge (ComParE). One of the sub-challenges is the COVID-19 cough sub-challenge (CCS) and COVID-19 speech sub-challenge (CSS). We only used data from CCS to merge with the previous datasets. The original dataset contains 725 recordings divided into Train (286 samples), Development (231 samples), and Test (208 samples). To enable comparison with previous studies, we used the same test partition. The other partitions (Train and Development) are merged with the previous datasets (Coswara and COUGHVID). The sampling rate was kept at 16 kHz, but the audio files were normalized in a range [−1, 1] following the same treatment on the previous datasets. This ComParE-CCS dataset was obtained by emailing the organizers of the challenge.

The summary of the number of samples for all datasets is shown in Table 2. Note that these numbers are after cough detection and segmentation processing blocks. There is a decrease in the number due to filtering by the cough detection method. For instance, the original ComParE-CCS train, development, and test sets are 286, 231, and 208 samples, respectively. After the cough detection and segmentation, the numbers of samples are 236, 134, and 154 samples, respectively. The same thing happens to the Coswara and COUGHVID datasets.

3.2 Acoustics features

Mel spectrogram is a powerful acoustic feature that is widely used in automatic speech recognition [34], language identification [35], speech emotion recognition [36], music tagging [37], and cough-based COVID-19 detection

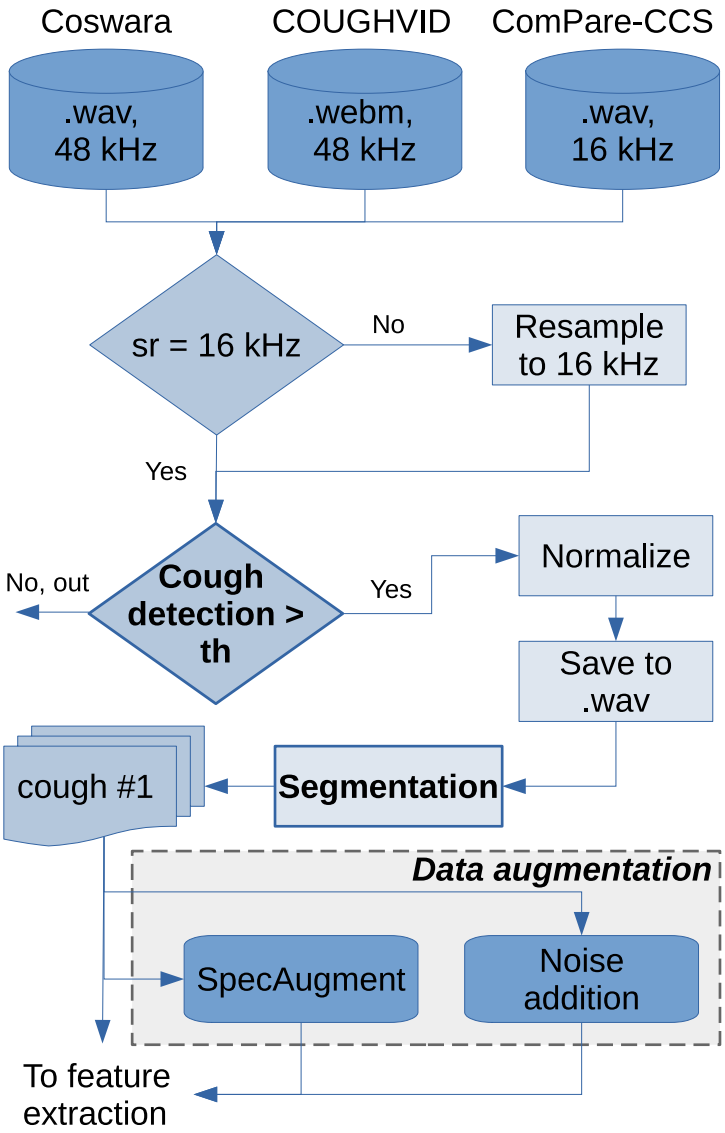

Fig. 2 The flowchart for the proposed method; data augmentation block is proposed to improve cough-based COVID-19 detection

Table 2 Number of samples for the final dataset (after cough detection and cough segmentation)

Dataset	Train	Devel	Test	Total
Coswara	564	100	–	664
COUGHVID	712	126	–	838
ComParE-CCS	236	134	154	524
Total	1512	360	154	2026

[21]. Borrowing the success of the mel spectrogram, we employed the log mel spectrogram as the acoustic feature for this study. The log mel spectrogram is computed by using the TorchAudio [38] toolkit. The parameters are set as follows: n_fft=1024, hop_length=320, n_mels=64, mel_fmin=0, and win_length=1024.

A Mel spectrogram is a spectrogram in the Mel scale. Suppose an audio spectrogram (Short Time Fourier Transform, STFT) with a center frequency f (Hertz, Hz; half of a sampling rate). The mel-scale conversion from Hertz is then given by,

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right). \quad (1)$$

The mel spectrogram's visualization is shown in Fig. 3, which is taken from the train data in the ComParE-CCS dataset.

Since the pretrained audio neural networks (PANNs) [39] is trained in log mel spectrogram, we convert the mel

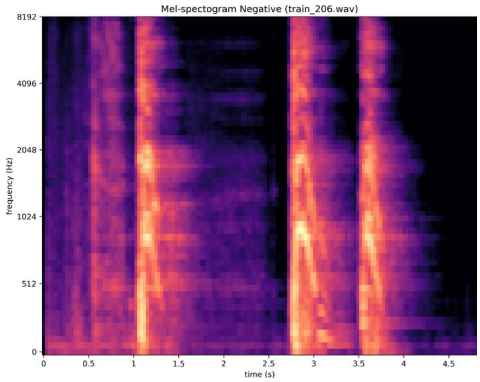

Fig. 3 Mel spectrogram of the cough sound

spectrogram to log mel spectrogram by using the following equation:

$$log\ S = 20 \times log_{10}\left(\frac{S}{ref}\right), \tag{2}$$

where S is the spectrogram in a mel scale (mel spectrogram) [eq. (1)] and *ref* is the reference value. The reference value in eq. (2) above is set to 1.0.

3.3 Transfer learning and classifiers

We employed a transfer learning from the previous pre-trained model for the AudioSet dataset (527 classes) [39]. The PANNS (pre-trained audio neural networks) outperformed the previous models on the original AudioSet task and other tasks (transfer learning), including ESC-50 (50 sound events), Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 Task 1 (Acoustic scene of 12 cities), DCASE 2028 Task 2 (automatic audio tagging of 41 labels), and GTZAN (music genre classification with 10 genres). We used PANNs with CNN14 architecture shown in Fig. 4. The model is trained with the AdamW optimizer (**Adam** with decoupled **W**eight decay regularization [40]), with a learning rate of 0.001, weight decay of 0.01, and batch size of 16. The model is trained for 100 epochs without early stopping.

The input of the classifier is a log mel spectrogram, and the output is binary classification (two classes) for positive and negative prediction. This output is a transfer learning from AudioSet with 527 classes to two COVID-19 classes via a linear layer (nodes = 2048). The pre-trained model is used as a feature extractor to obtain the embedding in the last layer (2048-dims). The output layer (in transfer learning) is activated by using a sigmoid function. The output is then either positive (patient, "1") or negative (control, "0").

3.4 Cough detection and segmentation

The goal of the cough detection processing block is to filter out non-cough signals and cough signals with low probabilities (low cough recognition rates). We employed a model from the previous study [29] by utilizing the Extreme Gradient Boosting (XGBoost) classifier [41]. The performance of the final model with threshold=0.8 is 88.1% of balanced accuracy (unweighted accuracy, in Table 3 [29]). The number of data is approximately 37,000 segments with cough labels "1" if cough is detected and "0" otherwise. There are 18 types of acoustic features for the cough detection model, from MFCC to spectral decrease. We analyze the distribution of each dataset (recognition rate of cough) and evaluate different thresholds of cough detection for cough-based COVID-19 classification in the experiments.

After cough detection, we also performed cough segmentation to split several coughs in a waveform into individual coughs. The segmentation is based on the silence region between coughs. Detail of the cough segmentation is described in [26], which is based on the previous studies on different cough segmentation methods [25, 29].

3.5 Data augmentation

Data augmentation is a technique to increase the size of the dataset by generating new samples from the existing samples. The goal of data augmentation is to increase the size of the dataset and reduce the overfitting problem. We employed the following data augmentation techniques: mixup [42], SpecAugment [43], and noise addition. The noise addition is performed by convolving the original audio data with MUSAN noise corpus [44]. The MUSAN dataset contains three types of noises: speech, music, and noise; we only chose noise type to add to the original audio data. The noise is added to the original audio data with a random SNR (signal-to-noise ratio) between 0 and 15 dB.

While mixup augmentation is used in the baseline, we evaluate the effectiveness of adding data with SpecAugment, noise addition, and a combination of both. A further evaluation for mixup augmentation is performed by evaluating different alpha mixup values ranging from 0.1 to 1.0 with a 0.1 step. This last evaluation is also performed to tune other hyperparameters: learning rate and weight decay.

3.6 Evaluation metric

We use a single metric, namely unweighted accuracy (UA), as the evaluation metric. The UA is defined as the number of correctly classified samples divided by the total number of samples per class. The UA is a common metric used in previous studies [23, 31]. Using this metric enables us to compare the performance of the proposed model with the

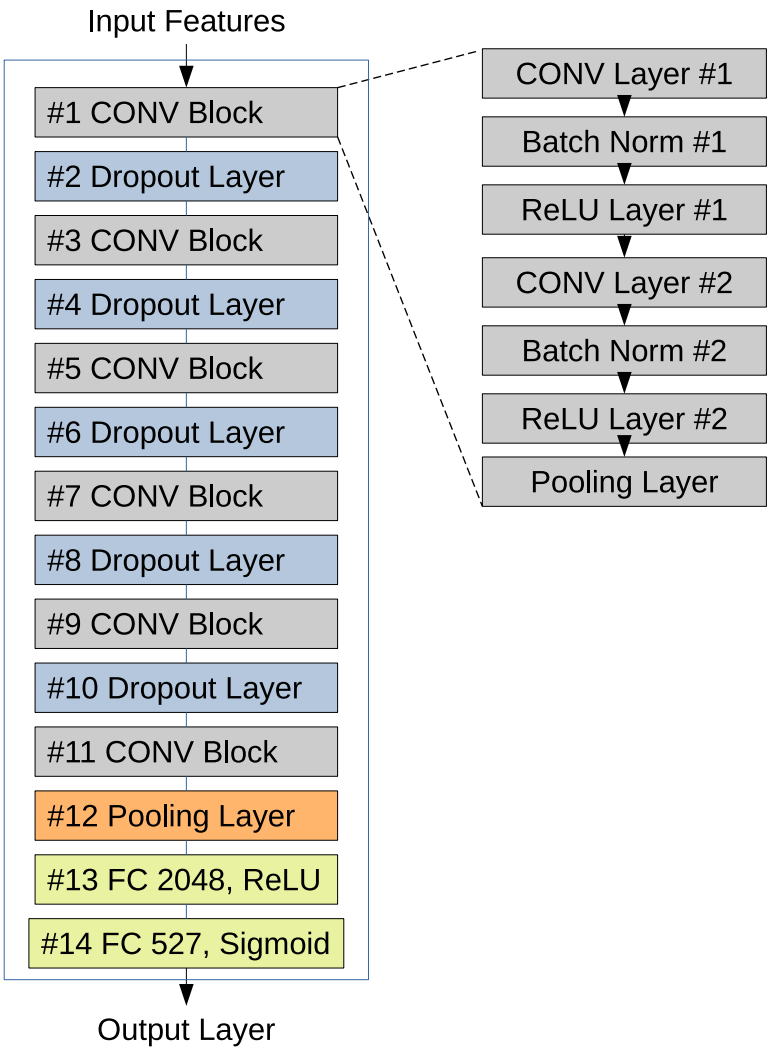

Fig. 4 Architecture of the CNN14 for transfer learning; the last embedding (2048-dims) is used as the input of transfer learning with a linear layer with 2048-node input and 2-node output

previous studies on the same test set (ComParE-CCS). This metric is also known as balanced accuracy or unweighted average recall (UAR). UA or UAR is formulated in eq. (3) below,

$$UA = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FN_i}, \tag{3}$$

where N is the number of classes, TP_i is the number of true positives for class i, and FN_i is the number of false negatives for class i. In this study, $N = 2$ (positive and negative classes).

4 Results and discussion

We present our results and discuss them in different parts: data exploration, cough detection, data augmentation, ablation study, and a summary of previous studies. These parts are described in the following subsections.

4.1 Data exploration

At first, it is necessary to explore the data to gain insights about information about them. We plot the duration of cough sounds in Fig. 5 to know the length of typical cough files. The majority of cough sounds have a duration of between 5-10 s. Another data exploration is the distribution between positive and negative samples. Figure 6 shows the distribution of positive and negative samples in the original three datasets. Naturally, the negative samples (healthy cough) are more than the positive samples (COVID-19 cough), as shown in Fig. 6. The use of negative cough may affect the classification of cough-based COVID-19 detection since some cough sounds in negative samples are not natural (i.e., artificial coughs).

Based on the previous study [26], we evaluate six different ratios for splitting training and validation data.

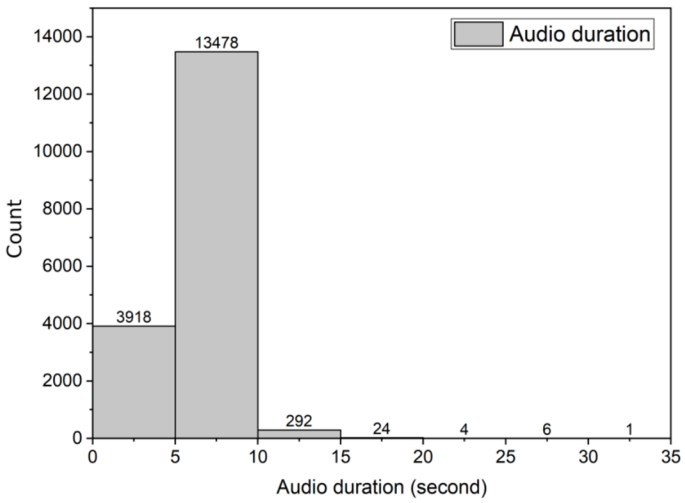

Fig. 5 Distribution of data based on duration

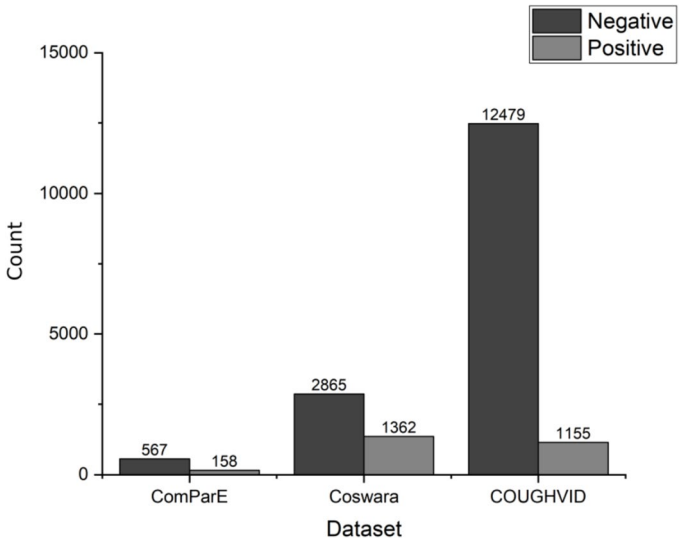

Fig. 6 Data distribution of positive and negative samples on the original three datasets

The different number of data for each split ratio is given in Fig. 7. The optimal value is given by a ratio of 85:15 for training:validation. Note that this ratio contains positive samples only for Coswara and COUGHVID datasets, according to our best experiments. For ComParE-CCS, the ratio includes negative samples that will be used for balancing other datasets. The next experiments are conducted using this 85:15 split ratio.

4.2 Effect of cough detection and segmentation

Our first proposal for enhancing the performance of cough-based COVID-19 detection is to detect cough sounds through specific thresholds. In this block, we want to filter cough sounds with a high recognition rate only (removing low-rate cough sounds, including non-cough sounds). The model used to detect cough is based on the XGB classifier [29]. We present the distribution of recognition rate for each dataset in Figs. 8, 9, and 10. It has been shown that there is a high number of data with a low recognition rate. For instance, in the COUGHVID dataset, there are 1015 data with a recognition rate below 10%. Removing the low recognition rate of cough sounds may improve the performance of cough-based COVID-19 detection.

We varied the thresholds to filter cough sounds from three datasets on values of 60%, 70%, 80%, and 90%. The distribution (data count) for each threshold is shown in Fig. 11. Table 3 shows the UA of each threshold. The results show that the best threshold is 90% for all three datasets with a UA of 75.54% (improved about 3% from the baseline without cough detection). It also shows that the smaller threshold leads to a lower UA, highlighting the importance of (our proposal by) filtering cough sounds with cough detection for COVID-19 detection. The next experiments are conducted using this 90% threshold cough detection.

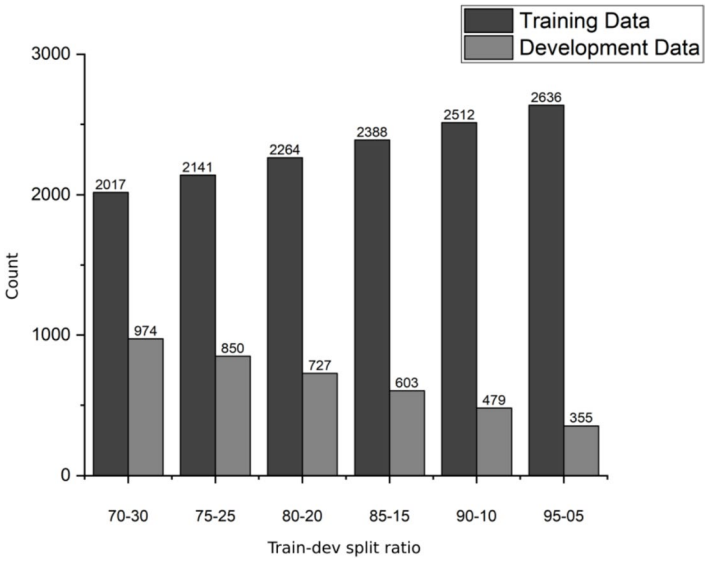

Fig. 7 Count of data on different train-test split ratio

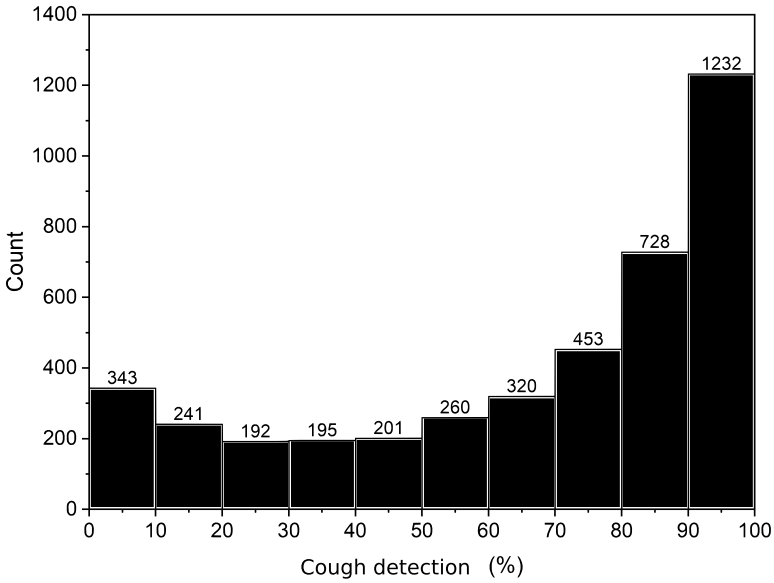

Fig. 8 Count of data for Coswara dataset on different cough detection thresholds

As reported in [26], the cough segmentation is also important for COVID-19 detection. The relative improvement (over cough detection) is 7.65% using hysteresis method. The next experiments are conducted using this 90% threshold cough detection and hysteresis comparator segmentation.

4.3 Effect of data augmentation

The last proposed processing block is data augmentation for enhancing COVID-19 detection by adding more data. Modern deep learning method, including CNN14, relies on large data since it tends to be more effective than small data [45, 46]. We evaluated three augmentation methods, i.e., (1) SpecAugment [43]; (2) noise addition using MUSAN noise corpus [44]; and (3) a combination of both (SpecAugment + Noise addition). The result is shown in Table 4. Note that all methods by default utilize mixup augmentation [42] in addition to these evaluated augmentation methods.

The results show that SpecAugment is the most effective method for this cough-based COVID-19 detection task. The UA of classification with SpecAugment is 86.39%; a large gap with the noise addition (UA = 81.4%) and a combination of both (UA = 83.19). It has also been found that adding more data does not always improve the performance of the model, as shown in other research [47]. With a double number of data, a combination of SpecAugment and noise addition only improves the UA by 1.5%, while a single SpecAugment method improves the UA by 4.7%. The next experiments are conducted to tune the hyperparameter based

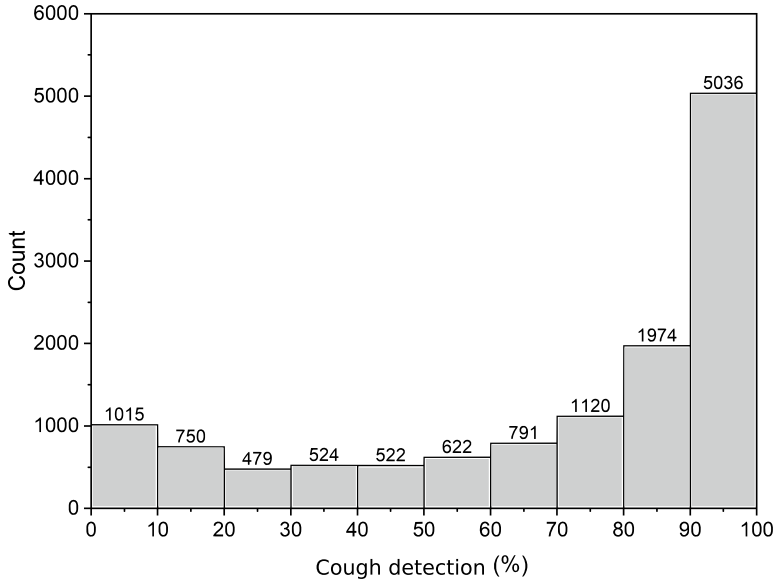

Fig. 9 Count of data for COUGHVID dataset on different cough detection thresholds

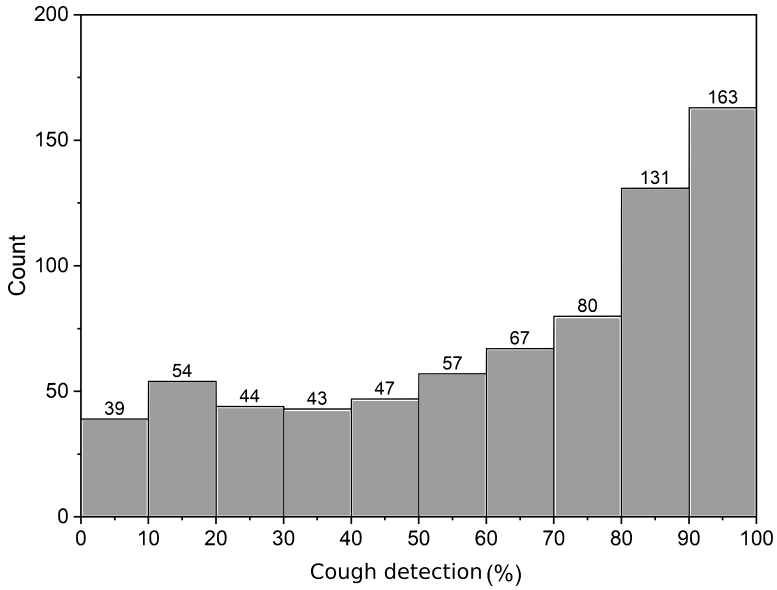

Fig. 10 Count of data for ComParE-CCS dataset on different cough detection thresholds

on this SpecAugment data augmentation method. Note that without augmentation, the result (along with other augmentation methods) still has a mixup augmentation method as a default configuration in the baseline method.

4.4 Ablation study

In the last part of this study, we conduct an ablation study by tuning three hyperparameters using fixed values in two parameters to find an optimal value for a single parameter in a specific range. This hyperparameter optimization (HPO) was done manually by changing a value in the optimized parameter while fixing two other parameters. These hyperparameters optimized linearly in this independent evaluation are alpha mixup, learning rate, and weight decay. Mixup is one of the data augmentation methods that is used to generate new data from existing data. The alpha mixup is a hyperparameter that controls the amount of mixup. The learning rate is a hyperparameter that controls the step size of the gradient descent. The weight decay is a hyperparameter that controls the regularization of the model. In this study, alpha mixup is varied in a range [0.1, 1.0], and learning rate and weight decay are varied in a range $[10^{-1}, 10^{-5}]$. Alpha mixup is varied with the addition of 10^{-1}. Both learning rate and weight decay are varied in a step of multiplication of 10^{-1}. The results are shown in Table 5.

We improve the performance of cough-based COVID-19 detection by changing the default alpha mixup [42] from 0.9

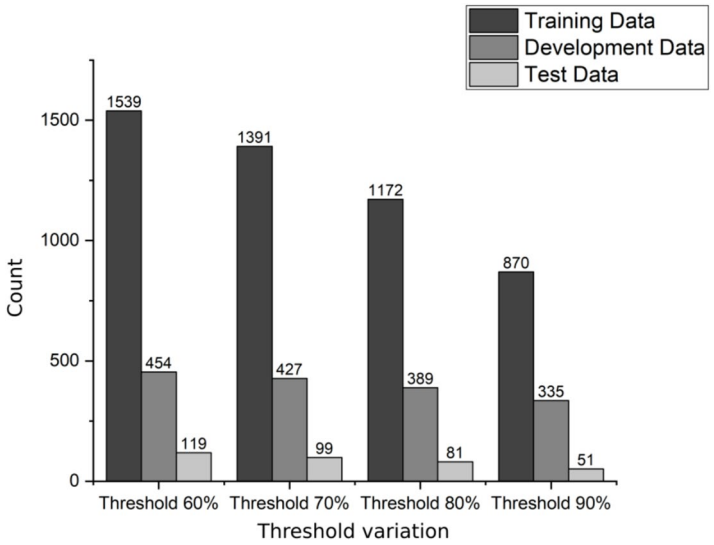

Fig. 11 Count of all data on different cough detection thresholds

Table 3 Data count results on different cough detection thresholds

Threshold (%)	Data count (# samples)
60	2201
70	2026
80	1769
90	1413

Table 4 UA results on different data augmentation methods

Segmentation method	Augmentation method	UA (%)
Hysteresis comparator	Without augmentation	81.68
	SpecAugment	**86.36**
	Noise addition	81.40
	SpecAugment + Noise addition	83.19

Bold shows the highest score

Table 5 UA results (%) on ablation study with a variation of alpha mixup, learning rate (LR), and weight decay (WD)

α	UA	LR	UA	WD	UA
0.1	86.50	0.1	53.53	0.1	87.87
0.2	86.91	0.01	49.17	**0.01**	**88.19**
0.3	87.74	**0.001**	**88.19**	0.001	84.43
0.4	85.12	0.0001	85.81	0.0001	86.36
0.5	**88.19**	0.00001	49.41	0.00001	87.46
0.6	84.84				
0.7	84.98				
0.8	86.36				
0.9	86.36				
1.0	85.81				

Bold shows the highest score

Default values are 0.9, 0.001, and 0.01, consequently

to 0.5. The default hyperparameters on previously reported UAs are 0.9, 0.001, and 0.0001 for the alpha mixup, learning rate, and weight decay, respectively. The other hyperparameters (learning rate and weight decay) are at their best. Choosing the proper parameter for mixup exhibits a potential solution to alleviate the issue of memorization and sensitivity during the training stage. The UA of this model, with 88.19%, is the highest obtained UA in this study.

4.5 Summary with the previous studies

Although it is not possible to compare our results directly with previous studies due to data differences (after cough detection and segmentation), we provide a summary study shown in Table 6. The results show that our proposed method outperforms the previous studies by a large margin, although we did not utilize the ensemble method. The most similar method to ours is the method by Cassanova et al. [23], from which our method is derived. Three processing blocks empirically improve that baseline method largely, including improvements from the original baseline methods [31]. A thorough summary without our result is also can be found in [48], in which the authors also discuss the result of speech-based COVID-19 detection in addition to the cough-based method.

5 Conclusion

In this paper, we evaluated three important processing blocks for training cough sounds for COVID-19 detection. These blocks are cough detection, cough segmentation and data augmentation. The gains using these blocks are 2.86%, 7.65% and 3.17%, respectively, relative to the addition of one block after another. These gains were

Table 6 Summary of UA scores (%) from different studies on the same ComParE-CCS 2021 test set; note that our method uses the same test set as others but with a different number of samples due to cough detection and segmentation methods

Reference	Data Aug	Feature	Classifier	Ensemble	UA
Baseline [31]	×	openSMILE, openXBOW, DeepSpctrum, AuDeep	SVM, End2You	√	73.90
Casanova et al. [23]	√	log mel spectrogram	CNN14	√	75.90
Illium et al. [49]	√	log mel spectrogram	Vision Transformer	×	76.90
Solera-Urena et al. [50]	×	TDNN-F, VGGish, PASE+	SVM	√	69.30
Suyanto et al. [26]	×	log mel spectrogram	CNN14	×	83.19
Ours	√	log mel spectrogram	CNN14+HPO	×	**88.19**

Bold shows the highest score

obtained using a cough detection rate at 90%, cough segmentation using hysteresis comparator method, and data augmentation using SpecAugment. Furthermore, we optimize the hyperparameters through an ablation study by tuning alpha mixup, learning rate, and weight decay. A proper value of alpha mixup (alpha=0.5) improves the UA 88.19%, which is the highest UA in this study. The learning rate and weight decay are at their best since the baseline method (learning rate=0.001 and weight decay = 0.01). The results show that the proposed method outperforms the previous studies by a large margin, although we did not utilize the ensemble method.

Future studies should be directed to test this research-stage method for preliminary COVID-19 detection in the real world. The results should be justified by accurate COVID-19 labeling, e.g., by PCR test. Explainable AI techniques can be used to explain and understand the model's decision, which is important for the clinical use of AI-based COVID-19 detection.

Acknowledgements B.T.A. and A.S. are supported by the New Energy and Industrial Technology Development Organization (NEDO) Japan Project No. JPNP20006 and JSPS KAKENHI Grant Number 24K02967. Z., S., and W. A. A. are supported by project number 1014/PKS/ITS/2022, funded by the Directorate of Research and Community Service, Sepuluh Nopember Institute of Technology (ITS), Indonesia. The authors would like to thank Dr. Dhany Arifianto of VibrasticLab ITS for allowing us to use his computational resources for this study.

Data availability The datasets generated and/or analyzed during the current study are obtained in the following schemes: Coswara: This dataset is available at the Coswara-Data repository: https://github.com/iiscleap/Coswara-Data. We used commit ID 401b516 during the study. COUGHVID: This dataset is available at the COUGHVID repository: https://zenodo.org/records/7024894. We used version 3.0 during the study. ComParE-CCS: This ComParE-CCS dataset is not publicly available. This is the dataset provided by the Computational Paralinguistic Challenge (ComParE) 2022 COVID-19 Cough sub-challenge organizer for their challenge. Please contact the authors [31] to obtain the dataset.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no Conflict of interest.

References

1. Hoda MN (2022) Editorial. Int J Inf Technol (Singapore) 14(7):3287–3290. https://doi.org/10.1007/s41870-022-01134-1
2. Yamin M (2020) Counting the cost of COVID-19. Int J Inf Technol (Singapore) 12(2):311–317. https://doi.org/10.1007/s41870-020-00466-0
3. Milling M, Pokorny FB, Bartl-Pokorny KD, Schuller BW (2022) Is speech the new blood? Recent progress in AI-based disease detection from audio in a nutshell. Front Digit Health 4(May):1–7. https://doi.org/10.3389/fdgth.2022.886615
4. Gupta R, Chaspari T, Kim J, Kumar N, Bone D, Narayanan S (2016) Pathological speech processing: State-of-the-art, current challenges, and future directions. In: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 6470–6474. IEEE, Shanghai . https://doi.org/10.1109/ICASSP.2016.7472923
5. Pramono RXA, Imtiaz SA, Rodriguez-Villegas E (2016) A cough-based algorithm for automatic diagnosis of pertussis. PLoS ONE 11(9):1–20. https://doi.org/10.1371/journal.pone.0162128
6. Al-khassaweneh M, Abdelrahman RB (2013) A signal processing approach for the diagnosis of asthma from cough sounds. J Med Eng Technol 37(3):165–171. https://doi.org/10.3109/03091902.2012.758322
7. Swarnkar V, Abeyratne UR, Chang AB, Amrulloh YA, Setyati A, Triasih R (2013) Automatic identification of wet and dry cough in pediatric patients with respiratory diseases. Ann Biomed Eng 41(5):1016–1028. https://doi.org/10.1007/s10439-013-0741-6
8. Bertini F, Allevi D, Lutero G, Calzà L, Montesi D (2021) An automatic Alzheimer's disease classifier based on spontaneous spoken English. Comput Speech Lang 72:101298. https://doi.org/10.1016/j.csl.2021.101298
9. Shuvo SB, Ali SN, Swapnil SI, Hasan T, Bhuiyan MIH (2021) A lightweight CNN model for detecting respiratory diseases from lung auscultation sounds using EMD-CWT-based hybrid scalogram. IEEE J Biomed Health Inf 25(7):2595–2603. https://doi.org/10.1109/JBHI.2020.3048006. arXiv:2009.04402
10. Han J, Xia T, Spathis D, Bondareva E, Brown C, Chauhan J, Dang T, Grammenos A, Hasthanasombat A, Floto A, Cicuta P, Mascolo C (2022) Sounds of COVID-19: exploring realistic performance of audio-based digital testing. Npj Digit Med 5(1):16. https://doi.org/10.1038/s41746-021-00553-x. arXiv:2106.15523
11. Anthes E (2020) Alexa, do I have COVID-19? Nature 586(7827):22–25. https://doi.org/10.1038/d41586-020-02732-4

12. Adjuik TA, Ananey-Obiri D (2022) Word2vec neural model-based techniqueto generate protein vectors for combating COVID-19: a machine learning approach. Int J Inf Technol (Singapore) 14(7):3291–3299. https://doi.org/10.1007/s41870-022-00949-2
13. Khanday AMUD, Rabani ST, Khan QR, Rouf N, Mohi Ud Din M (2020) Machine learning based approaches for detecting COVID-19 using clinical text data. Int J Inf Technol (Singapore) 12(3):731–739. https://doi.org/10.1007/s41870-020-00495-9
14. Singh D, Singh BK, Behera AK (2023) A real-time correlation model between lung sounds & clinical data for asthmatic patients. Int J Inf Technol 15(1):39–44. https://doi.org/10.1007/s41870-022-01138-x
15. Quatieri TF, Talkar T, Palmer JS (2020) A framework for biomarkers of COVID-19 based on coordination of speech-production subsystems. IEEE Open J Eng Med Biol 1:203–206. https://doi.org/10.1109/OJEMB.2020.2998051
16. Islam R, Abdel-Raheem E, Tarique M (2022) A study of using cough sounds and deep neural networks for the early detection of Covid-19. Biomed Eng Adv 3:100025. https://doi.org/10.1016/j.bea.2022.100025
17. Vahedian-azimi A, Keramatfar A, Asiaee M, Atashi SS, Nourbakhsh M (2021) Do you have COVID-19? An artificial intelligence-based screening tool for COVID-19 using acoustic parameters. J Acoust Soc Am 150(3):1945–1953. https://doi.org/10.1121/10.0006104
18. Bartl-Pokorny KD, Pokorny FB, Batliner A, Amiriparian S, Semertzidou A, Eyben F, Kramer E, Schmidt F, Schönweiler R, Wehler M, Schuller BW (2021) The voice of COVID-19: acoustic correlates of infection in sustained vowels. J Acoust Soc Am 149(6):4377–4383. https://doi.org/10.1121/10.0005194
19. Hamidi M, Zealouk O, Satori H, Laaidi N, Salek A (2023) COVID-19 assessment using HMM cough recognition system. Int J Inf Technol 15(1):193–201. https://doi.org/10.1007/s41870-022-01120-7
20. Hasan I, Dhawan P, Rizvi SAM, Dhir S (2023) Data analytics and knowledge management approach for COVID-19 prediction and control. Int J Inf Technol (Singapore) 15(2):937–954. https://doi.org/10.1007/s41870-022-00967-0
21. Mohammed EA, Keyhani M, Sanati-Nezhad A, Hejazi SH, Far BH (2021) An ensemble learning approach to digital corona virus preliminary screening from cough sounds. Sci Rep 11(1):1–11. https://doi.org/10.1038/s41598-021-95042-2
22. Chowdhury NK, Kabir MA, Rahman MM, Islam SMS (2022) Machine learning for detecting COVID-19 from cough sounds: an ensemble-based MCDM method. Comput Biol Med 145(March):105405. https://doi.org/10.1016/j.compbiomed.2022.105405
23. Casanova E, Candido Jr, A, Fernandes Jr, RC, Finger M, Gris LRS, Ponti MA, Pinto da Silva DP (2021) Transfer learning and data augmentation techniques to the COVID-19 identification tasks in ComParE 2021. In: Interspeech 2021, pp. 446–450. ISCA, ISCA https://doi.org/10.21437/Interspeech.2021-1798. https://www.isca-speech.org/archive/interspeech_2021/casanova21_interspeech.html
24. Sharma G, Umapathy K, Krishnan S (2022) Audio texture analysis of COVID-19 cough, breath, and speech sounds. Biomed Signal Process Control. https://doi.org/10.1016/j.bspc.2022.103703
25. Atmaja BT, Zanjabila Suyanto Sasou A (2023) Comparing hysteresis comparator and RMS threshold methods for automatic single cough segmentations. Int J Inf Technol. https://doi.org/10.1007/s41870-023-01626-8
26. Suyanto Z, Atmaja BT, Asmoro WA (2024) Performance improvement of Covid-19 cough detection based on deep learning with segmentation methods. J Appl Data Sci 5(2):520–531
27. Wang CC, Pan CA, Hung JW (2008) Silence feature normalization for robust speech recognition in additive noise environments. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, pp 1028–1031
28. Atmaja BT, Akagi M (2020) The effect of silence feature in dimensional speech emotion recognition. In: 10th international conference on speech prosody 2020, pp 26–30. ISCA, Tokyo. https://doi.org/10.21437/SpeechProsody.2020-6
29. Orlandic L, Teijeiro T, Atienza D (2021) The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. Sci Data 8(1):156. https://doi.org/10.1038/s41597-021-00937-4
30. Haritaoglu ED, Rasmussen N, Tan DCH, J, JR, Xiao J, Chaudhari G, Rajput A, Govindan P, Canham C, Chen W, Yamaura M, Gomezjurado L, Broukhim A, Khanzada A, Pilanci M (2022) Using deep learning with large aggregated datasets for COVID-19 classification from cough, 1–10 arXiv:2201.01669
31. Schuller BW, Batliner A, Bergler C, Mascolo C, Han J, Lefter I, Kaya H, Amiriparian S, Baird A, Stappen L, Ottl S, Gerczuk M, Tzirakis P, Brown C, Chauhan J, Grammenos A, Hasthanasombat A, Spathis D, Xia T, Cicuta P, Rothkrantz LJM, Zwerts JA, Treep J, Kaandorp CS (2021) The INTERSPEECH 2021 computational paralinguistics challenge: COVID-19 cough, COVID-19 speech, escalation & primates. In: Interspeech 2021, pp 431–435. ISCA, ISCA . https://doi.org/10.21437/Interspeech.2021-19
32. Sharma N, Krishnan P, Kumar R, Ramoji S, Chetupalli SR, Nirmala R, Kumar Ghosh P, Ganapathy S (2020) Coswara - A database of breathing, cough, and voice sounds for COVID-19 diagnosis. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH 2020-Octob, 4811–4815 arXiv:2005.10548. https://doi.org/10.21437/Interspeech.2020-2768
33. McFee B, Lostanlen V, McVicar M, Metsai A, Balke S, Thomé C, Raffel C, Malek A, Lee D, Zalkow F, Lee K, Nieto O, Mason J, Ellis D, Yamamoto R, Seyfarth S, Battenberg E, Morozov V, Bittner R, Choi K, Moore J, Wei Z, Hidaka S, Nullmightybofo Friesch P, Stöter F-R, Hereñú D, Kim T, Vollrath M, Weiss A (2020) librosa/librosa: 0.7.2 . https://doi.org/10.5281/ZENODO.3606573
34. Guo J, Sainath TN, Weiss RJ (2019) A Spelling Correction Model for End-to-end Speech Recognition. In: ICASSP 2019 - 2019 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 5651–5655. IEEE, Brighton, UK. https://doi.org/10.1109/ICASSP.2019.8683745
35. Choi K, Wang Y (2021) Listen, Read, and Identify: Multimodal Singing Language Identification. In: Proc Ofthe 22nd Int Society for Music Information Retrieval Conf, pp 121–127
36. Liu Z-T, Xiao P, Li D-Y, Hao M (2019) Speaker-Independent Speech Emotion Recognition Based on CNN-BLSTM and Multiple SVMs. In: International conference on intelligent robotics and applications, pp 481–491
37. Choi K, Fazekas G, Sandler M, Cho K (2018) A comparison of audio signal preprocessing methods for deep neural networks on music tagging. In: 2018 26th European signal processing conference (EUSIPCO), pp 1870–1874. IEEE, Rome, Italy. https://doi.org/10.23919/EUSIPCO.2018.8553106
38. Yang Y-Y, Hira M, Ni Z, Astafurov A, Chen C, Puhrsch C, Pollack D, Genzel D, Greenberg D, Yang EZ, Lian J, Hwang J, Chen J, Goldsborough P, Narenthiran S, Watanabe S, Chintala S, Quenneville-Belair V (2022) Torchaudio: building blocks for audio and speech processing. In: ICASSP 2022 - 2022 IEEE international conference on acoustics, speech and signal processing (ICASSP), vol 2022-May, pp 6982–6986. IEEE, https://doi.org/10.1109/ICASSP43922.2022.9747236
39. Kong Q, Cao Y, Iqbal T, Wang Y, Wang W, Plumbley MD (2020) PANNs: large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Trans Audio Speech

Lang Process 28(1):2880–2894. https://doi.org/10.1109/TASLP.2020.3030497. arXiv:1912.10211
40. Loshchilov I, Hutter F (2017) Decoupled weight decay regularization. In: 7th International conference on learning representations, ICLR arXiv:1711.05101
41. Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system arXiv:1603.02754. https://doi.org/10.1145/2939672.2939785
42. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D (2018) MixUp: beyond empirical risk minimization. In: 6th international conference on learning representations, ICLR 2018 - Conference Track Proceedings, pp 1–13
43. Park DS, Chan W, Zhang Y, Chiu CC, Zoph B, Cubuk ED, Le QV (2019) Specaugment: a simple data augmentation method for automatic speech recognition. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, vol 2019-Septe, pp 2613–2617. https://doi.org/10.21437/Interspeech.2019-2680
44. Snyder D, Chen G, Povey D (2015) MUSAN: a music, speech, and noise corpus arXiv:1510.08484
45. Halevy A, Norvig P, Pereira F (2009) The unreasonable effectiveness of data. IEEE Intell Syst 24(2):8–12. https://doi.org/10.1109/MIS.2009.36
46. Goodfellow I, Bengio Y, Courville A (2015) Deep Learning Book. MIT Press, Cambridge
47. Atmaja BT, Sasou A (2022) Effects of data augmentations on speech emotion recognition. Sensors 22(16):5941. https://doi.org/10.3390/s22165941
48. Coppock H, Akman A, Bergler C, Gerczuk M, Brown C, Chauhan J, Grammenos A, Hasthanasombat A, Spathis D, Xia T, Cicuta P, Han J, Amiriparian S, Baird A, Stappen L, Ottl S, Tzirakis P, Batliner A, Mascolo C, Schuller BW (2023) A summary of the ComParE COVID-19 challenges. Front Digit Health 5:1–2. https://doi.org/10.3389/fdgth.2023.1058163. arXiv:2202.08981
49. Illium S, Müller R, Sedlmeier A, Popien CL (2021) Visual transformers for primates classification and covid detection. Proc Ann Conf Int Speech Commun Assoc INTERSPEECH 6:4341–4345. https://doi.org/10.21437/Interspeech.2021-273
50. Solera-Ureña R, Botelho C, Teixeira F, Rolland T, Abad A, Trancoso I (2021) Transfer learning-based cough representations for automatic detection of COVID-19. Proc Ann Conf Int Speech Commun Assoc INTERSPEECH 6:4336–4340. https://doi.org/10.21437/Interspeech.2021-1702