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

Computer Speech and Language 73 (2022) 101320

Contents lists available at ScienceDirect

Computer Speech & Language

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

Towards sound based testing of COVID-19—Summary of the first Diagnostics of COVID-19 using Acoustics (DiCOVA) Challenge

Neeraj Kumar Sharma[1], Ananya Muguli[1], Prashant Krishnan[1], Rohit Kumar, Srikanth Raj Chetupalli, Sriram Ganapathy *

Learning and Extraction of Acoustic Patterns (LEAP) Lab, Electrical Engineering, Indian Institute of Science, Bangalore, India

ARTICLE INFO

Keywords:
COVID-19
Acoustics
Machine learning
Respiratory diagnosis
Healthcare

ABSTRACT

The technology development for point-of-care tests (POCTs) targeting respiratory diseases has witnessed a growing demand in the recent past. Investigating the presence of acoustic biomarkers in modalities such as cough, breathing and speech sounds, and using them for building POCTs can offer fast, contactless and inexpensive testing. In view of this, over the past year, we launched the "Coswara" project to collect cough, breathing and speech sound recordings via worldwide crowdsourcing. With this data, a call for development of diagnostic tools was announced in the Interspeech 2021 as a special session titled "Diagnostics of COVID-19 using Acoustics (DiCOVA) Challenge". The goal was to bring together researchers and practitioners interested in developing acoustics-based COVID-19 POCTs by enabling them to work on the same set of development and test datasets. As part of the challenge, datasets with breathing, cough, and speech sound samples from COVID-19 and non-COVID-19 individuals were released to the participants. The challenge consisted of two tracks. The Track-1 focused only on cough sounds, and participants competed in a leaderboard setting. In Track-2, breathing and speech samples were provided for the participants, without a competitive leaderboard. The challenge attracted 85 plus registrations with 29 final submissions for Track-1. This paper describes the challenge (datasets, tasks, baseline system), and presents a focused summary of the various systems submitted by the participating teams. An analysis of the results from the top four teams showed that a fusion of the scores from these teams yields an area-under-the-receiver operating curve (AUC-ROC) of 95.1% on the blind test data. By summarizing the lessons learned, we foresee the challenge overview in this paper to help accelerate technological development of acoustic-based POCTs.

1. Introduction

The viral respiratory infection caused by the novel coronavirus, SARS-CoV-2, termed as the coronavirus disease 2019 (COVID-19), was declared a pandemic by the World Health Organization (WHO) in March 2020. The current understanding of COVID-19 prognosis suggests that the virus infects the nasopharynx and then spreads to the lower respiratory tract (Schaefer et al., 2020). One of the key strategies to combat the rapid spread of infection across populations is to perform rapid and large-scale testing.

Currently, the prominent COVID-19 testing methodologies take a molecular sensing approach. The gold-standard technique, termed as reverse transcription polymerase chain reaction (RT-PCR) (Corman et al., 2020), relies on using nasopharyngeal or throat

* Corresponding author.
E-mail address: sriramg@iisc.ac.in (S. Ganapathy).
[1] Equal contribution.

https://doi.org/10.1016/j.csl.2021.101320
Received 18 June 2021; Received in revised form 20 September 2021; Accepted 28 October 2021
Available online 24 November 2021

Table 1
A list of publicly accessible COVID-19 audio datasets.

Ref	Dataset	Sound categories	Access	COVID/non-COVID samples[a]	Method
Orlandic et al. (2021)	COUGHVID	Cough	Public	1155/27 550	Crowdsourced
Sharma et al. (2020)	Coswara	Cough, speech, breathing	Public	345/1785	Crowdsourced
Virufy COVID-19 Open Cough Dataset (2021)	Virufy	Cough	Public	7/9	Hospital
Cohen-McFarlane et al. (2020)	NoCoCoDa	Cough	On request	13/NA	YouTube
Brown et al. (2020)	COVID-19 sounds	Cough, breathing	On request	141/318	Crowdsourced

[a]Samples refers to count of distinct audio records from human subjects. Each audio record is composed of a set of sound recordings corresponding to the stated sound categories.

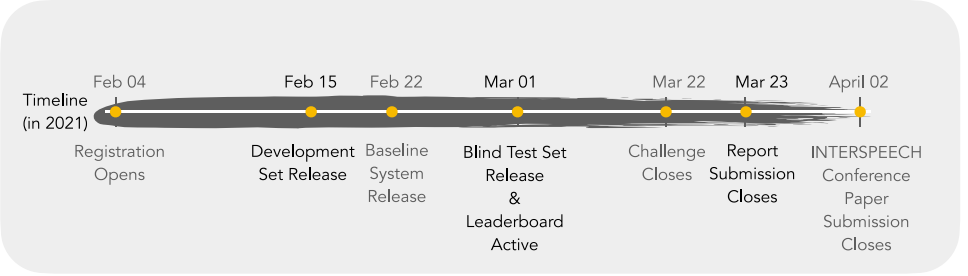

Fig. 1. The DiCOVA challenge timeline.

swab samples. The swab sample is treated with chemical reagents enabling isolation of the ribonucleic acid (RNA), followed by deoxyribonucleic acid (DNA) formation, amplification and analysis, facilitating the detection of COVID-19 genome in the sample. However, this approach has several limitations. The swab sample collection procedure violates physical distancing (Target product profiles, 2020). The processing of these samples requires a well equipped laboratory, with readily available chemical reagents and expert analysts. Further, the turnaround time for test results can vary from several hours to a few days. The protein based rapid antigen testing (RAT) (Peeling et al., 2021) improves over the speed of detection while being inferior to the RT-PCR in detection performance. The RAT test also involves the need for chemical reagents.

In view of the above mentioned limitations in molecular testing approaches (namely, RT-PCR and RAT), there is a need to design highly specific, rapid and easy-to-use point-of-care tests (POCTs) that could identify the infected individuals in a decentralized manner. Using acoustics for developing such a POCT would overcome various limitations in terms of speed, and cost, and also allow scalable remote testing.

1.1. Exploring acoustics based testing

The use of acoustics for diagnosis of pertussis (Pramono et al., 2016), tuberculosis (Botha et al., 2018), childhood pneumonia (Abeyratne et al., 2013), and asthma (Hee et al., 2019) has been explored using cough sounds recorded with portable devices. As COVID-19 is an infection affecting the respiratory pathways (Li et al., 2021), recently, researchers have made efforts towards COVID-19 acoustic data collection. A list of acoustic datasets is provided in Table 1. Building on these datasets, few studies have evaluated the possibility of COVID-19 detection using acoustics. Brown et al. (2020) used cough and breathing sounds jointly and attempted a binary classification task of separating COVID-19 infected individuals from healthy. The dataset was collected through crowd-sourcing, and the analysis was done on 141 COVID-19 infected individuals. The authors reported a performance between 80–82% AUC-ROC (area-under-the-receiver operating characteristic curve). Agbley et al. (2020) demonstrated 81% specificity (at 43% sensitivity) on a subset of the COUGHVID dataset (Orlandic et al., 2021). Imran et al. (2020) studied cough sound samples from four groups of individuals, namely, healthy, and those with bronchitis, pertussis, and COVID-19 infection. They report an accuracy of 92.6%. Laguarta et al. (2020) used a large sample set of COVID-19 infected individuals and report an AUC-ROC performance of 97.0%. Andreu-Perez et al. (2021) create a controlled dataset by collecting cough sound samples from patients visiting hospitals, and they report 98.8% AUC-ROC.

Although these studies are encouraging, they suffer from some limitations. They do not use a common dataset, and a few are based on privately collected datasets. The ratio of COVID-19 patients to healthy (or non-COVID) is different in every study. The performance metrics are also different across studies. Some of the studies report performance per-cough bout, and others report per-patient. Further, most of the studies have not bench-marked on other open source datasets, making it difficult to compare among the various propositions.

1.2. Contribution

We launched the "Diagnostics of COVID-19 using Acoustics (DiCOVA) Challenge" (Muguli et al., 2021) with two primary goals. Firstly, to encourage the speech and audio researchers to analyze acoustics of cough and speech sounds for a problem of immediate

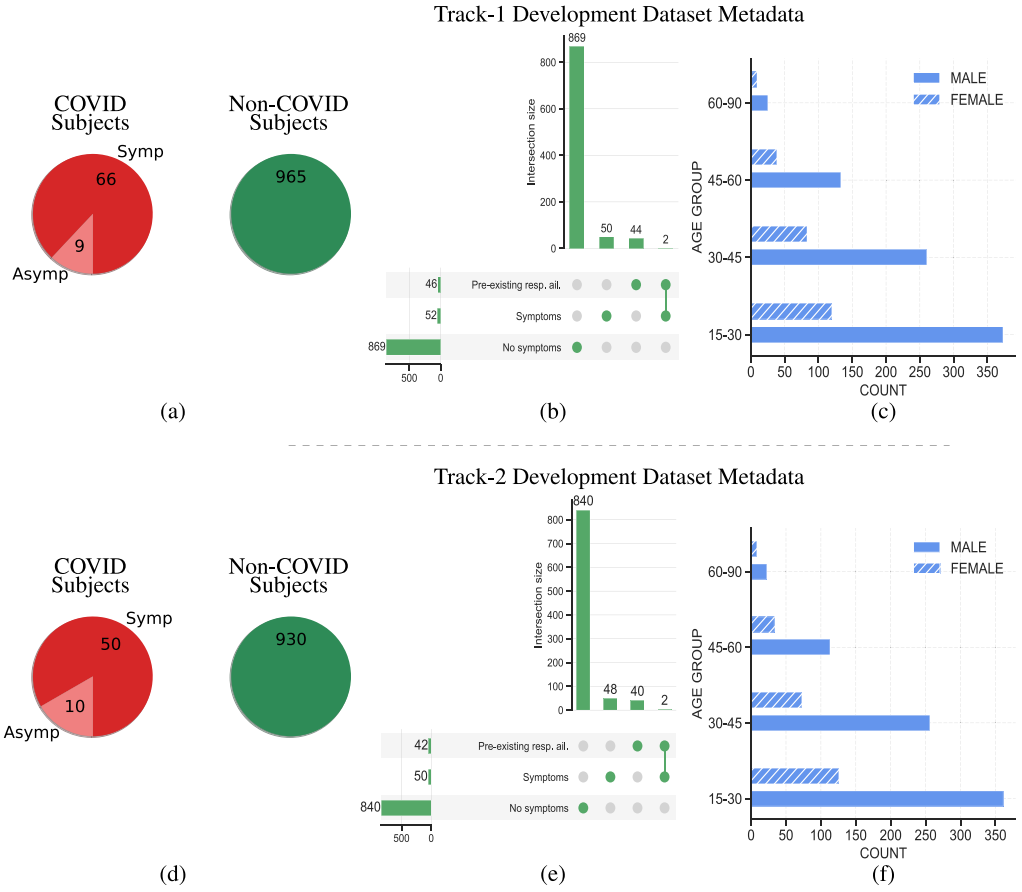

Fig. 2. An illustration of Track-1 and Track-2 development datasets. Here, (a,d) show the COVID and non-COVID pool size in terms of number of individuals; (b,e) show the breakdown of non-COVID individuals into categories of no symptoms, symptoms (cold, cough), and pre-existing respiratory ailment (asthma, chronic lung disease, pneumonia); (c,f) depicts the age group distribution in the development dataset.

societal relevance. The challenge was launched under the umbrella of Interspeech 2021, and participants were given an option to submit their findings to a special session in this flagship conference. Secondly, and more importantly, to provide a benchmark for monitoring the progress in acoustic based diagnostics of COVID-19. The development and (blind) test datasets were provided to the participants to facilitate design and evaluation of classifier systems. A leaderboard was created allowing participants to rank order their performance against others. This paper describes the details of the challenge including the dataset, the baseline system (Section 2), and provides a summary of the various submitted systems (Section 3). An analysis of the scores submitted by the top teams (Section 4), and the insights gained from the challenge are also presented (Section 6).

2. DiCOVA Challenge

The DiCOVA challenge[2] was launched on 04−Feb, 2021 and the challenge lasted till 23−Mar, 2021. The participation was through a registration process. A development set, a baseline system, and a blind test set was provided to all registered participants. A timeline of the challenge is shown in Fig. 1. A remote server based scoring system with a leaderboard setting was created. This provided near real-time ranking and monitoring progress of each team on the blind test set.[3] The call for participation in the challenge attracted 85 plus registrations. Out of this, 29 teams made final submissions on the blind test set.

[2] https://dicova2021.github.io/.

[3] https://competitions.codalab.org/competitions/29640.

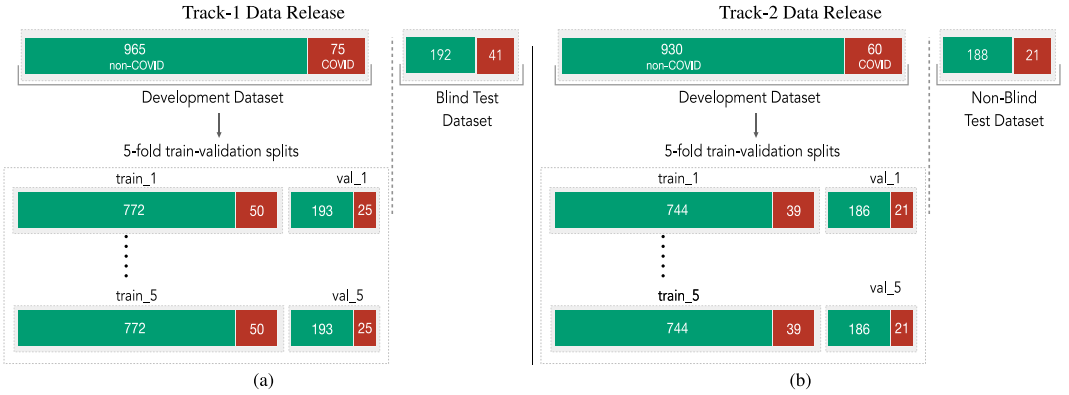

Fig. 3. Illustration of dataset splits for Track-1 (cough) and Track-2 (breathing and speech).

2.1. Dataset

The challenge dataset is derived from the Coswara dataset (Sharma et al., 2020), a crowd-sourced dataset of sound recordings. The Coswara data is collected using a website.[4] The volunteers from across the globe, age groups and health conditions were requested to record their sound data in a quiet environment using an internet connected device (like, mobile phone or computer).

The participants initially provide demographic information like age and gender. An account of their current health status in the form of a questionnaire of symptoms as well as pre-existing conditions like respiratory ailments and co-morbidity are recorded. The web based tool also records the result of the COVID-19 test conducted and the possibility of exposure to the virus through primary contacts.

The acoustic data from each subject contains 9 audio categories, namely, (*a*) shallow and deep breathing (2 types), (*b*) shallow and heavy cough (2 types), (*c*) sustained phonation of vowels [æ] (as in bat), [i] (as in beet), and [u] (as in boot) (3 types), and (*d*) fast and normal pace number counting (2 types). The dataset collection protocol was approved by the Human Ethics Committee of the Indian Institute of Science, Bangalore, and P. D. Hinduja National Hospital and Medical Research Center, Mumbai, India.

The DiCOVA Challenge used a subset of the Coswara dataset, sampled from the data collected between April-2020 and Feb-2021. The sampling included only age group of 15–90 years. The subjects with health status of "recovered" (who were COVID positive however fully recovered from the infection) and "exposed" (suspecting exposure to the virus) were not included in the dataset. Further, subjects with audio recordings of duration less than 500 ms were discarded. The resulting curated subject pool was divided into the following two groups.

- **non-COVID**: Subjects self reported as healthy, having symptoms such as cold/cough or having pre-existing respiratory ailments (like asthma, pneumonia, chronic lung disease) but were not tested positive for COVID-19.
- **COVID**: Subjects self-declared as COVID-19 positive (asymptomatic or symptomatic with mild/moderate infection)

The DiCOVA 2021 challenge featured two tracks. The Track-1 development dataset composed of (heavy) cough sound recordings from 1040 subjects. The Track-2 development dataset composed of deep breathing, vowel [i], and number counting (normal pace) speech recordings from 990 subjects. An illustration of the important metadata details in the development set is provided in Fig. 2. About 70% of the subjects were male. The majority of the participants lie in the age group of 15–40 years. Also, the dataset is highly imbalanced with less than 10% of the participants belonging to the COVID category. We retained this class imbalance in the challenge as this reflects the typical real-world scenario.

In the data release, the development dataset was further divided into train and validation splits. The splits are illustrated in Fig. 3. A meta-analysis study of COVID-19 symptoms by Li et al. (2021) found cough (53.9%) as a common symptoms in 281,641 COVID-19 infected individuals. In addition, prior efforts on data collection and modeling largely focused on the cough samples (see Table 1) (Orlandic et al., 2021; Brown et al., 2020). Owing to this, the challenge emphasized progress in Track-1. A leaderboard was created and the participants competed by uploading their scores for a blind test dataset and monitoring the performance. The Track-2 featured the test dataset, without any leaderboard-style competition and encouraged the participants to carry out an exploratory analysis.

2.2. Audio specifications

The crowd-sourced dataset reflects a good representation of real-world data with sensor variability arising from diverse recording devices. For the challenge, we re-sampled all audio recordings to 44.1 kHz and compressed them to FLAC (Free Lossless Audio

[4] https://coswara.iisc.ac.in/.

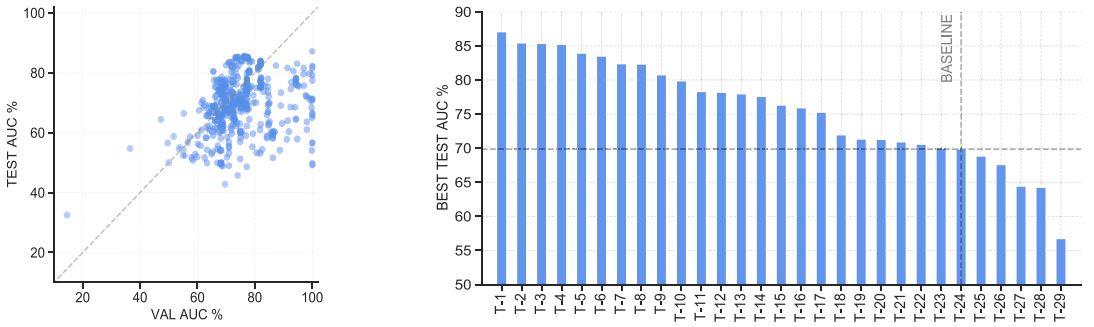

Fig. 4. (a) A scatter plot of the average five-fold validation AUC versus test AUC performance for every submission on the leaderboard. (b) Test set AUC performance in rank ordered manner for each of the system submissions. Here, AUC refers to AUC-ROC.

Codec) format for ease of distribution. The average duration of Track-1 development set cough recordings is 4.72 (standard deviation ($S.D.$) $\pm$ 2.07) s. The average duration of Track-2 development set audio recordings is -breathing 17.65 ($\pm$6.20) s, vowel [i] 12.37 ($\pm$6.12) s, and number counting speech 14.78 ($\pm$3.87) s.

2.3. Task

- **Track 1**: The task focused on cough audio samples only. This was the primary track of the challenge with most teams participating only in this track. A leaderboard website was hosted for the challenge enabling teams to evaluate their system performance (validation set and blind test set). The participating teams were required to submit the COVID probability score for each audio file in the validation and test sets. The leaderboard website computed the ROC-AUC and the specificity/sensitivity. Every team was provided a maximum of 25 tickets for submitting scores to the leaderboard.
- **Track 2**: Track-2 explored the use of recordings other than cough for the task of COVID diagnostics. The audio recordings released in this track composed of breathing, speech related to sustained phonation of vowel [i] and number counting (1–20). The development and (non-blind) test sets were released concurrently, without any formal leaderboard style evaluation and competition.

The data and the baseline system setup were provided to the registered teams after signing a terms and conditions document. As per the document, the teams were not allowed to use the publicly available Coswara dataset.[5]

2.4. Evaluation metrics

The focus of the challenge was binary classification, that is, detecting COVID or non-COVID using acoustics. As the dataset was imbalanced, we choose not to use accuracy as an evaluation metric. Each team submitted COVID probability scores ($\in [0, 1]$, a higher value indicating a higher likelihood of COVID infection) for the list of validation/test audio recordings. For performance evaluation, we used the scores with the ground truth labels to compute the receiver operating characteristics (ROC) curve. The curve was obtained by varying the decision threshold between 0–1 with a step size of 0.0001. The area under the resulting ROC curve, AUC-ROC, was used as a performance measure for the classifier, where the area was computed using the trapezoidal method. The AUC-ROC formed the primary evaluation metric. Further, specificity (true negative rate), at a sensitivity (true positive rate) greater than or equal to 80% was used as a secondary evaluation metric. For brevity, we will refer to AUC-ROC by AUC in the rest of the paper.

2.5. Baseline system

The baseline system was implemented using tools from the `scikit-learn` Python library (Pedregosa et al., 2011).

Pre-processing: For every audio file, the signal was normalized in amplitude. Using a sound activity detection threshold of 0.01 and a buffer size of 50 ms on either side of a sample, any region of the audio signal with amplitude lower than threshold was discarded. Also, initial and final 20 ms snippets of the audio were removed to avoid abrupt start and end activity in the recordings.

Feature extraction: The baseline system used the 13 dimensional mel-frequency cepstral coefficients (MFCCs), its delta and delta–delta coefficients, computed over 1024 samples (23.2 ms), with a hop of 441 samples (10 ms). The resulting feature dimension was 39×1.

Classifiers: The following three classifiers were designed.

[5] https://github.com/iiscleap/Coswara-Data.

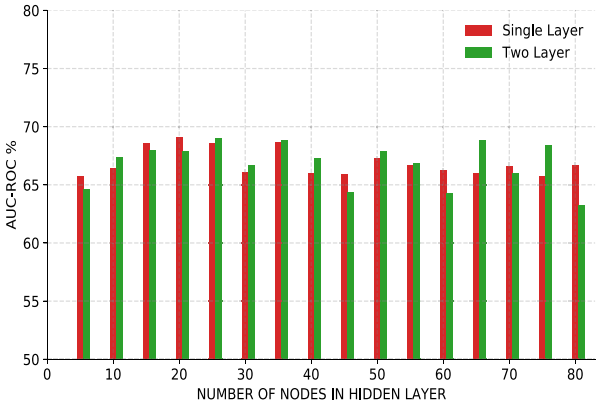

Fig. 5. Variation in AUC as a function of number of nodes in the hidden layer of an MLP classifier. The AUC-ROC% is the average over the five validation folds.

- Logistic Regression (LR): The LR classifier was trained for 25 epochs. The binary cross entropy (BCE) loss with a l_2 regularization strength of 0.01 was used for optimizing the model.
- Multi layer perceptron (MLP): A single-layer perceptron model with 25 hidden units and `tanh()` activation was used. Similar to the LR model, the BCE loss with a ℓ_2 regularization of strength 0.001 was optimized for parameter estimation. The loss was optimized using Adam optimizer with an initial learning rate of 0.001. The COVID samples were over-sampled to compensate for the data imbalance (weighted BCE loss).
- Random Forests (RF): A random forest classifier was trained with 50 trees using *Gini* impurity criterion for tree growing.

In the weighted BCE loss used in LR and MLP, the class errors are weighted. That is,

$$\mathbb{L} = \sum_{i=0}^{N-1} -w_0 y_i \log_{10}(\hat{y}_i) - w_1(1-y_i)\log_{10}(1-\hat{y}_i) \tag{1}$$

where $\mathbb{L}$ is the loss, N is the number of training samples, y_i and $\hat{y}_i$ are the true and predicted labels, and w_0 and w_1 are the weights associated with non-COVID and COVID classes. We choose w_0 and w_1 as the inverse of fraction of samples associated with the corresponding class in the training set. In RF, class weights are used to weigh the Gini impurity criterion for finding splits for tree growing. In the terminal nodes of each tree, class weights are again taken into consideration for class prediction via "weighted majority" vote (Chen et al., 2004).

In the MLP design, the goal was to develop a shallow architecture for the baseline system and encourage participants to build deep networks with pre-training from other datasets. Our internal analysis had shown over-fitting issues for deep architectures when trained only with the challenge data. Hence, a single hidden layer architecture with `tanh()` was chosen. Ablation experiments were carried out to decide on the number of nodes in the hidden layer of this MLP. A grid search in the range of 5–80 nodes in steps of 5 showed average AUC-ROC (over the five validation folds) in the range of 63–70%. The AUC-ROC improved with addition of nodes from 5 to 25 and after this the increase did not so a monotonic improvement. This is shown in Fig. 5. Also, shown is the AUC-ROC obtained using a two hidden layer MLP. The performance is in the similar range as that of single layer MLP. For the baseline system we opted for a single layer MLP with 25 hidden units.

Inference and performance: To obtain a classification score for an audio recording: (*i*) the file was pre-processed, (*ii*) frame-level MFCC features were extracted, (*iii*) frame-level probability scores were computed using the trained model(s), and (*iv*) all the frame scores were averaged to obtain a single COVID probability score for the audio recording. For evaluation on the test set files, the probability scores from five validation models (for a classifier type) were averaged to obtain the final score.

Table 2 depicts the performance of the three classifiers on the validation folds and the test sets. All classifiers performed better than chance. For Track-1, the AUC for the test set was better for the MLP classifier (69.85% AUC). For Track-2, RF gave the best AUC in all sound categories (65.27% – 76.85% AUC).

Further, among the category of acoustic sounds, the breathing samples provided the best AUC (76.85%) performance followed by vowel sound [i] (75.47%). The baseline system code[6] was provided to the participants as a reference for setting up a classifier training and scoring pipeline.

[6] https://github.com/dicova2021/Track-1-baseline.

Table 2
The baseline system performance on Track-1 and Track-2 on development set (5-fold val) and test set.

Track	Sound	Model	Performance (AUC%)	
			Val. (std. dev)	Test
1	Cough	LR	66.95 (±3.89)	61.97
		MLP	68.54 (±3.69)	**69.85**
		RF	70.69 (±3.10)	67.59
2	Breathing	LR	60.95 (±4.85)	60.94
		MLP	72.47 (±4.38)	71.52
		RF	75.17 (±2.75)	**76.85**
	Vowel [i]	LR	71.48 (±1.23)	67.71
		MLP	70.39 (±4.11)	73.19
		RF	69.73 (±4.31)	**75.47**
	Speech	LR	68.93 (±2.44)	61.22
		MLP	73.57 (±1.59)	61.13
		RF	69.61 (±3.49)	**65.27**

Table 3
Summary of submitted systems in terms of feature and model configurations. The specificity (%) is reported at a sensitivity of 80%.

Team ID rank-wise	Implementation				Performance	
	Data aug.	Features	Classifiers	Ensemble	Test AUC%	Test Spec.%
T-1 (Mahanta et al., 2021)	✓	MFCCs	CNN	✗	87.07	83.33
T-2 (Chang et al., 2021)	✓	mel-spectrogram	ResNet50	✓	85.43	82.29
T-3 (Harvill et al., 2021)[a]	✓	mel-spectrogram	LSTM	✓	85.35	71.88
T-4 (Södergren et al., 2021)[a]	✗	openSMILE	RF, SVM	✓	85.21	81.25
T-5 (Singh et al., 2021)	✓	MFCC	CNN, LSTM ResNet	✓	83.93	70.83
T-6 (Das et al., 2021)[a]	✗	ERB-spectrogram	RF, MLP	✓	83.49	77.08
T-7 (Elizalde and Tompkins, 2021)	✗	YAMNet, OpenL3	Extra Trees	✗	82.37	72.92
T-8[b]	✗	mel-spectrogram	ResNet34	✗	82.32	67.19
T-9 (Avila et al., 2021)[a]	✓	openSMILE	SVM, CNN	✓	80.75	63.54
T-10[b]	✓	mel-spectrogram	ResNet34	✓	79.86	68.23
T-11[b]	✓	mel-spectrogram	VGG13	✓	78.30	50.52
T-12 (Karas and Schuller, 2021)[a]	✗	openSMILE, embeddings	SVM, LSTM	✓	78.18	58.85
T-13[b]	✗	openSMILE, mel-spectrogram	DNN, VGGCNN	✓	77.96	59.38
T-14[b]	✗	handcrafted, log mel-spectrogram	MLP	✗	77.58	60.42
T-15 (Kamble et al., 2021)[a]	✗	TECC + $\Delta + \Delta\Delta$	LightGBM	✗	76.31	53.65
T-16 (Banerjee and Nilhani, 2021)	✓	mel-Spectrograms	ResNet50	✗	75.91	62.50
T-17[b]	✗	MFCCs	CNN	✗	75.26	55.21
T-18 (Bhosale et al., 2021)[a]	✗	MFCCs	DNN	✗	71.94	47.40
T-21 (Mallol-Ragolta et al., 2021)[a]	✗	mel-Spectrogram Images	ResNet18	✗	70.91	47.40
T-24 (Muguli et al., 2021)[a]	✗	MFCCs	MLP	✗	69.85	53.65
T-27 (Deshpande and Schuller, 2021)[a]	✗	handcrafted features	LSTM	✗	64.42	40.10

[a]Denotes report accepted in Interspeech 2021.
[b]Denotes team did not give consent for the public release of the report.

3. Track-1: Submitted systems overview

A total of 28 teams (plus the baseline system) participated in the Track-1 leaderboard. Out of these, 20 teams submitted their system reports describing the explored approaches.[7] In this section, we provide a brief overview of the submissions, emphasizing on the obtained performances and explored classifiers, features, model ensembling and data augmentation techniques.

3.1. Performance

In total, 23 out of the 28 teams reported a performance better than the baseline system. We refer to the teams with Team IDs corresponding to their rank on the leaderboard, that is, best AUC performance as T-1 and so on. A performance summary of all the submitted systems on the validation and the blind test data is given in Fig. 4. Fig. 4(a) depicts a comparison of the validation and test results. Interestingly, there is a slight positive correlation between test and validation performance. For some teams, the validation performances exceed 95% AUC. Deducing from the system reports, these performances are primarily due to training on the whole development dataset without removing the validation data. Fig. 4(b) depicts the best AUC posted by 29 participating teams (including baseline) on the blind test data. The best AUC performance on the test data was 87.07%, a significant improvement over the baseline AUC (that is, 69.85%).

[7] The system reports are available at https://dicova2021.github.io/#reports.

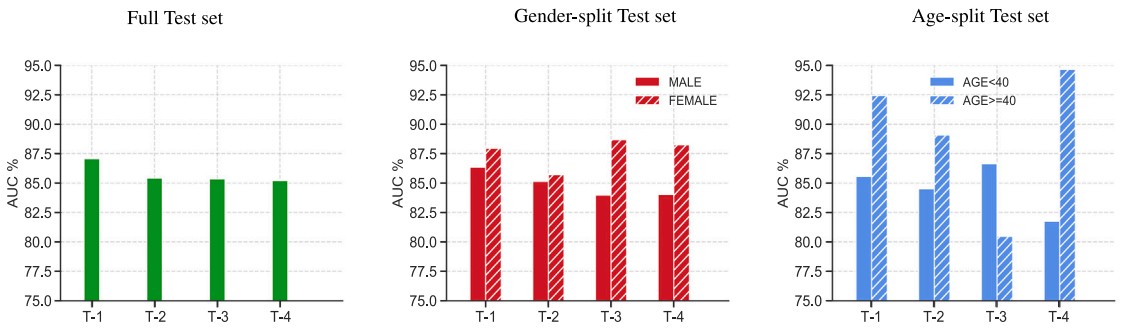

Fig. 6. Illustration of AUC performance on full test set, and test set split by gender, and age. For male set: 171 subjects (27 COVID), for female set: 62 subjects (14 COVID), for age < 40 set: 180 subjects (31 COVID), and for age ≥ 40 set: 53 subjects (10 COVID).

3.2. Features

The teams designed and experimented with a wide spectrum of features. A concise highlight is provided in Table 3. A majority of the teams used mel-spectrograms, mel-frequency cepstral coefficients (Davis and Mermelstein, 1980), or equivalent rectangular bandwidth (ERB) (Smith and Abel, 1999) spectrograms (15 submissions out 21). Further, the openSMILE features (Eyben et al., 2010), which consist of statistical measures extracted on low-level acoustic feature descriptors, were explored by 4 teams. Few teams explored features derived using Teager energy based cepstral coefficients (TECC Kamble and Patil, 2019; T-15), and pool of short-term features such as short-term energy, zero-crossing rate, and voicing (T-5, T-14, T-27). Other teams resorted to using embeddings derived from pre-trained neural networks as features. These included VGGish (Hershey et al., 2017), DeepSpectrum (Amiriparian et al., 2017), OpenL3 (Cramer et al., 2019), YAMNet (Plakal and Ellis, 2020) embeddings (T-7, T-12), and x-vectors (Snyder et al., 2018) (T-15).

3.3. Classifiers

The teams explored various classifier models (see Table 3). These included classical machine learning models, such as decision trees, random forests (RFs), and support vector machines (SVMs), and modern deep learning models, such as convolutional neural networks (CNNs), long short-term memory (LSTM) networks, and residual networks (ResNet). Several teams also attempted an ensemble of models to improve the final system performance.

The CNNs were explored by teams (T-1, T-5, T-10, T-17). Variants of CNNs with residual connections and recording level average pooling to deal with the variable length input were developed by teams (T-2, T-5, T-8, T-9, T-10, T-13, T-16, T-21). Citing the improved ability of LSTMs to handle variable length inputs, (T-3, T-5, T-12, T-27) explored these models. The classical ML approaches of random forest, logistic regression and SVMs were used by (T-4, T-6, T-12, T-18). LightGBM (Gradient Boosting Machine) (Ke et al., 2017) model was explored by (T-15), and extra trees classifiers were studied by (T-7). Pre-training was also studied in several systems (T-3, T-17). Autoencoder style pre-training was used by (T-17). Several teams had also experimented with transfer learning from architectures pre-trained on image based models (T-2, T-8, T-13) and audio based models (T-10, T-13).

3.4. Model ensembling

The fusion of scores from different classifier architectures was explored by multiple teams (T-3, T-4, T-6 T-10, T-11, T-12, T-13). The fusion of multiple features was explored by (T-13). Further, (T-2, T-3) investigated score fusion of outputs obtained from the model tuned on the five validation folds.

3.5. Data augmentation

Data augmentation is a popular strategy in which external or synthetic audio data is used in training of deep network models. Five teams reported using this strategy by including COUGHVID cough dataset (Orlandic et al., 2021) (publicly available), adding Gaussian noise at varying SNRs, or doing audio manipulations (pitch shifting, time-scaling, etc., via tools such as audiomentations[8]). Few teams also used data augmentation approaches to circumvent the problem of class imbalance. These included T-1 using mixup (Zhang et al., 2017), (T-3, T-9, T-11) using SpecAugment (Park et al., 2019), (T-2, T-5, T-9) using additive noise, T-21 using sample replication, and T-5 using Vocal-Tract Length Perturbation (VTLP) (Jaitly and Hinton, 2013), to increase the sample counts of the minority class.

[8] https://github.com/iver56/audiomentations.

Table 4
A comparison of AUC and sensitivity of top four teams, their score fusion and the baseline system.

Performance measures	Team T-1	Team T-2	Team T-3	Team T-4	Fusion	Baseline
AUC %	**87.07**	85.43	85.35	85.21	95.07	69.85
Sensitivity (at 95% Specificity)	46.34	39.02	**60.97**	29.27	70.73	17.07

Besides these, other strategies for training included gender aware training (T-21), using focal loss (Lin et al., 2017) objective function (T-2, T-8, T-11), and hyper-parameter tuning using model searching algorithm TPOT (T-7) (Le et al., 2019).

In the next section, we discuss in detail the approaches used by the Track-1 four top performing teams.

4. Track-1: Top performers

4.1. T-1: The brogrammers

The team (Mahanta et al., 2021) focused on using a multi-layered CNN network architecture. Special emphasis was laid on having a small number of learnable parameters. Every audio segment was trimmed or zero padded to 7 s. For feature extraction, this segment was represented using 15 dimensional MFCC features per frame, and a matrix of 15×302 frames was obtained. A cascade of a CNN and fully connected layers, with max-pooling and ReLU non-linearities, was used in the neural network architecture. For data augmentation, the team used the audiomentations tool. The classifier was trained using binary cross entropy (BCE) loss to output COVID probability score. The team did not report performing any system combination unlike several other participating teams.

4.2. T-2: NUS-Mornin system

The team focused (Chang et al., 2021) on using the residual network (ResNet) model with spectrogram images as features. To overcome the limitations of data scarcity and imbalance, the team resorted to three key strategies. Firstly, data augmentation was done by adding Gaussian noise to spectrograms. Secondly, focal loss function was used instead of binary cross entropy loss. Thirdly, the ResNet50 was pre-trained on ImageNet followed by fine-tuning on DiCOVA development set and an ensemble of four models was used to generate final COVID probability scores.

4.3. T-3: UIUC SST system

The team (Harvill et al., 2021) used long short term memory (LSTM) models. With the motivation of generative modeling of mel-spectrogram for capturing informative features of cough, the team proposed using the auto-regressive predictive coding (APC) (Oord et al., 2018). The APC is used to pre-train the initial LSTM layers operating on the input mel-spectrogram. The additional layers of the full network, which was composed of BLSTM and fully connected layers, was trained using the DiCOVA development set. As the number of model parameters was high, the team also used data augmentation using COUGHVID dataset (Orlandic et al., 2021) and SpecAugment (Park et al., 2019) tool. The binary cross entropy was chosen as the loss function. The final COVID-19 probability score was obtained as an average of several similar models, trained on development data subsets or sampled at different checkpoints during training.

4.4. T-4: The North system

The team (Södergren et al., 2021) explored classical machine learning models like random forests (RF), support vector machines (SVM), and multi-layer perceptron (MLP) rather than deep learning models. The features used were the 6373 dimensional openSMILE functional features (Eyben et al., 2010). The openSMILE features were z-score normalized to prevent feature domination. The hyper-parameters of the models were tuned to obtain the best results. The SVM models alone provided an AUC of 85.1% on the test data. The RF and the MLP scored an AUC of 82.15 and 75.65, respectively. The final scores were obtained by a weighted average of the probability scores from the RF and SVM models, with weights of 0.25 and 0.75, respectively.

4.5. Top 4 teams: Fairness

Here, we present a fairness analysis of the scores generated by the top 4 teams. We particularly focus on gender-wise and age-wise performance on the test set. Fig. 6 depicts this performance. Interestingly, all the four teams gave a better performance for female subjects. Similarly, the test dataset was divided into two groups based on subjects with age < 40 and age ≥ 40. Here, the top two teams had a considerably higher AUC for age ≥ 40 subjects, while T-3 had a lower AUC for this age group and T-4 had the highest. In summary, the performance of top four teams did not reflect the bias in the development data (70% male subjects, and largely in age ≤ 40 group; see Fig. 2).

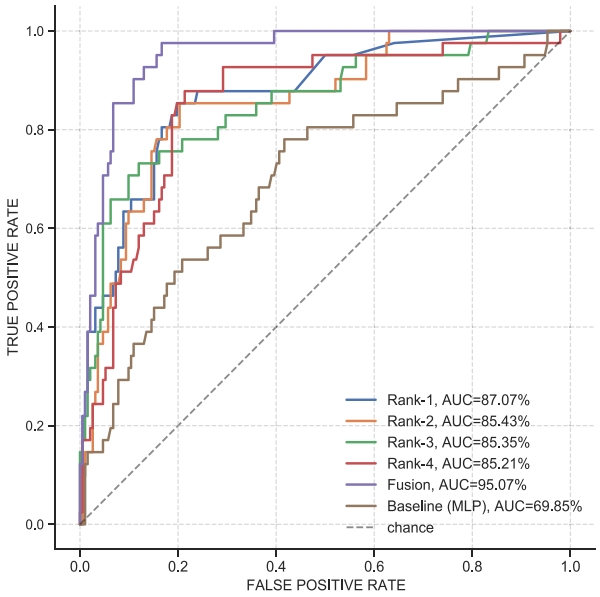

Fig. 7. Illustration of ROCs obtained on the test set for the top four teams. The ROC associated with the hypothetical score fusion system obtained using the top four teams is also shown.

4.6. Top 4 teams: Score fusion

The systems from the top four teams differ in terms of features, model architectures, and data augmentation strategies. We consider a simple arithmetic mean fusion of the scores from top 4 teams. Let p_{ij}, $1 \leq i \leq N$ and $1 \leq j \leq T$, be the COVID probability score predicted by the jth team submission for the ith subject in the test data. Here, N denotes the number of subjects in the test set, and T, denoting the number of top teams, is four. The scores are first calibrated by correcting for the range as follows.

$$\hat{p}_{ij} = \frac{p_{ij} - p_{min,j}}{p_{max,j} - p_{min,j}}, \tag{2}$$

where $p_{min,j} = \min\left(p_{1j}, \ldots, p_{Nj}\right)$ and $p_{max,j} = \max\left(p_{1j}, \ldots, p_{Nj}\right)$. The fused scores are obtained as,

$$p_{i,f} = \frac{1}{4}\sum_{j=1}^{4} \hat{p}_{ij}. \tag{3}$$

The ROC obtained using these prediction scores is denoted by *Fusion* in Fig. 7. The Fusion system ROC gives an AUC of 95.10%, a significant improvement over each of the individual system results. Table 4 depicts the sensitivity of the top four systems, the fusion, and baseline (MLP) at 95% specificity. The fused model surpasses all the other models and achieves a sensitivity of 70.7%.

5. Track-2: Systems overview

This track was an exploratory track. It did not feature a leaderboard and did not require system report submission to the organizers. Hence, we have a summary of explorations carried by two teams only (only two reports were available with the details on Track-2 submission).

Ritwik et al. (2021) performed a detailed analysis on the COVID detection performance obtained with different spectral features for each of the three sound categories, namely, breathing, vowel [i] and counting. The authors explored acoustical features included MFCCs, Gaussian mixture model (GMM) based super vectors, formant frequency features, fundamental frequency values and its harmonics. The study suggests formant frequency features as the best performing feature for the binary task. Further, complimentary information is present in the three sound categories. A fusion of probability scores from each sound category, for each subject, gave an AUC 73.4% on validation folds and 71.7% AUC on the test (same as eval) dataset.

Deshpande and Schuller (2021) explored the estimates of breathing patterns, obtained from different sound categories, for COVID-19 detection. Towards this, an encoder which predicts breathing pattern from speech signals was designed using a subset of UCL Speech Breath Monitoring (UCL-SBM) database (Schuller et al., 2020). This pre-trained encoder is then used to predict the breathing patterns from breathing, vowel-[i], and counting sound categories, separately. The estimated breathing patterns are then used as feature vectors to train a decoder model to predict COVID-19 status. Interestingly, the breathing features performed superior to MFCCs for vowel-e and counting sound categories. Further, a combination of breathing and MFCCs features performed better than either one of these features. Across the three sound categories provided in Track-2, the average validation AUC ranged between 62.0–66.0% and the test AUC ranged between 61.0–67.0%.

6. Discussion

6.1. Challenge accomplishments

The challenge problem statement for Track-1 required the design of a binary classifier. A clear problem statement, with a well-defined evaluation metric (AUC), encouraged a significant number of registrations. This included 85 plus teams from around the globe, with a good representation from both industry and academia. The 28 teams which completed the challenge came from 9 different countries. Additionally, 8 teams associated themselves with industry. Among the submissions, 23 out of the 28 teams exhibited a performance well above the baseline system AUC (see Fig. 4(b)).

Altogether, the challenge provided a platform for researchers to explore a healthcare problem of immense and timely societal impact. The results indicate potential in using acoustics for COVID-19 POCT development. The challenge turnaround time was 49 days, and the progress made by different teams in this short time span highlighted their efforts. Eleven studies pursued in this challenge (Muguli et al., 2021; Das et al., 2021; Mallol-Ragolta et al., 2021; Ritwik et al., 2021; Deshpande and Schuller, 2021; Karas and Schuller, 2021; Bhosale et al., 2021; Södergren et al., 2021; Harvill et al., 2021; Kamble et al., 2021; Avila et al., 2021), after going through the peer review process, were presented at the DiCOVA Special Session, Interspeech 2021 Conference (on 31 Aug 2021).

The World Health Organization (WHO) has stated that a sensitivity $\geq 80\%$ (at a specificity $\geq 97\%$) is necessary for an acceptable POCT tool (Target product profiles, 2020). The top four teams fell short of this benchmark (see Table 4), indicating that there is scope for further development in future. Interestingly, a simple combination of the scores from the systems of these teams achieves a performance more closer to this benchmark. This suggests some ways to reap advantage via collaboration between multiple teams for improved tool development. The development of such an acoustic based diagnostic tool for COVID-19 diagnosis would offer multiple advantages in terms of speed, cost, portability, and accuracy.

6.2. Limitations and future scope

The challenge, being first of its kind, also had its own limitations. We discuss some of these below.

- The development dataset had class imbalance, with a majority of the samples belonging to the non-COVID class. Although the imbalance reflects the prevalence of the infection in the population, it will be ideal to improve the balance in future challenges. The Coswara dataset (Sharma et al., 2020) developed by our team is being regularly updated with more samples. As of August 2021, it contains data from close to 345 COVID-19 positive individuals and 1785 non-COVID individuals.
- A majority of the DiCOVA dataset samples came from India. While the cultural dependence of cough and breathing is not well established, it will be ideal to evaluate the performance on datasets collected from multiple geographical sites. Towards this, future challenges can include demographically balanced datasets, with close collaborations between multiple sites involved in the data collection efforts.
- The task involved in the challenge simplified to a binary classification setting. However, in a practical scenario, there are multiple respiratory ailments resulting from bacterial, fungal, or viral infections, with each condition potentially leaving a unique bio-marker. The future challenges can target multi-class categorization. This will also widen the usability of the tool.
- The data did not contain information regarding the progression of the disease (or the time elapsed since the positive COVID-19 test). Also, the subjects in the "recovered from COVID-19" and "exposed to COVID-19 patient" categories were excluded in the challenge dataset. The leaderboard and system highlights reported were limited to the cough recordings only. As seen in Table 2, analysis using breathing and speech signals can also yield performance results comparable to those observed in cough recordings. In addition, the Coswara tool (Sharma et al., 2020) also records the symptom data from participants. Using a combination of various sound categories and symptoms in developing a COVID-19 detection tool might further push the detection performance.
- In the DiCOVA challenge, the performance ranking of the teams was based on AUC-ROC metric. This conveyed the model's ability to perform binary classification of COVID and non-COVID subjects. However, the challenge did not emphasize model interpretability and explainability as key requirements. In a healthcare scenario, the interpretability of the model decisions may be as important as the accuracy. Hence, future challenges should encourage this aspect. For example, a recent work by Xia et al. (2021) proposes an ensemble framework for quantifying decision uncertainty. Multiple classification models are developed and a disagreement across the learned model during testing phase is used as a measure of uncertainty in decision. In future, it is also important to focus on reproducibility of the models, and lower memory and computational foot-prints as these will benefit design and deployment of tool in mobile devices.

Recently, on 12 Aug 2021, we launched the Second DiCOVA Challenge[9] which attempts to circumvent some of the above limitations. It features three sound categories, namely, breathing, cough and speech, and a leaderboard for each category. In a separate track, the participants are also encouraged to fuse scores or decisions from classifiers designed on multiple sound categories. Further, in comparison to the first DiCOVA Challenge, the dataset is larger in size.

[9] https://dicovachallenge.github.io/.

Declaration of competing interest

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

Acknowledgments

The authors would like to thank the Verisk Analytics, Inc. for funding related to challenge organization and system development. The authors would also like to thank the Department of Science and Technology (DST), Government of India, for providing financial support for the Coswara Project through the RAKSHAK programme. The authors would like to thank the Organizing Committee, Interspeech 2021 for hosting this challenge under the umbrella of International Speech Communication Association (ISCA). The authors would like to express gratitude to Anand Mohan for the design of the web based Coswara data collection platform, Dr. Prasanta Kumar Ghosh for coordination in challenge organization, and Dr. Nirmala R., Dr. Shrirama Bhat, Dr. Lancelot Pinto, and Dr. Viral Nanda for their coordination in Coswara data collection.

References

Abeyratne, U.R., Swarnkar, V., Setyati, A., Triasih, R., 2013. Cough sound analysis can rapidly diagnose childhood pneumonia. Ann. Biomed. Eng. 41 (11), 2448–2462.

Agbley, B.L.Y., Li, J., Haq, A., Cobbinah, B., Kulevome, D., Agbefu, P.A., Eleeza, B., 2020. Wavelet-based cough signal decomposition for multimodal classification. In: 17th Intl. Computer Conference on Wavelet Active Media Technology and Information Processing). IEEE, pp. 5–9.

Amiriparian, S., Gerczuk, M., Ottl, S., Cummins, N., Freitag, M., Pugachevskiy, S., Baird, A., Schuller, B., 2017. Snore sound classification using image-based deep spectrum features. In: Proc. Interspeech, pp. 3512–3516.

Andreu-Perez, J., Perez-Espinosa, H., Timonet, E., Kiani, M., Giron-Perez, M.I., Benitez-Trinidad, A.B., Jarchi, D., Rosales, A., Gkatzoulis, N., Reyes-Galaviz, O.F., Torres, A., Alberto Reyes-Garcia, C., Ali, Z., Rivas, F., 2021. A generic deep learning based cough analysis system from clinically validated samples for point-of-need COVID-19 test and severity levels. IEEE Trans. Serv. Comput. 18, 1.

Avila, F., Poorjam, A.H., Mittal, D., Dognin, C., Muguli, A., Kumar, R., Chetupalli, S.R., Ganapathy, S., Singh, M., 2021. Investigating feature selection and explainability for COVID-19 diagnostics from cough sounds. In: Proc. Interspeech 2021. pp. 951–955. http://dx.doi.org/10.21437/Interspeech.2021-2197.

Banerjee, A., Nilhani, A., 2021. A residual network based deep learning model for detection of COVID-19 from cough sounds. arXiv:2106.02348.

Bhosale, S., Tiwari, U., Chakraborty, R., Kopparapu, S.K., 2021. Contrastive learning of cough descriptors for automatic COVID-19 preliminary diagnosis. In: Proc. Interspeech 2021. pp. 946–950. http://dx.doi.org/10.21437/Interspeech.2021-1249.

Botha, G., Theron, G., Warren, R., Klopper, M., Dheda, K., Van Helden, P., Niesler, T., 2018. Detection of tuberculosis by automatic cough sound analysis. Physiol. Meas. 39 (4), 045005.

Brown, C., Chauhan, J., Grammenos, A., Han, J., Hasthanasombat, A., Spathis, D., Xia, T., Cicuta, P., Mascolo, C., 2020. Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data. In: Proc. 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Association for Computing Machinery, New York, NY, USA, pp. 3474–3484.

Chang, J., Cui, S., Feng, M., 2021. DiCOVA-Net: Diagnosing COVID-19 using Acoustics based on Deep Residual Network for the DiCOVA Challenge 2021. Tech. Rep., DiCOVA Challenge, URL https://dicova2021.github.io/docs/reports/team_Jemery_DiCOVA_2021_Challenge_System_Report.pdf.

Chen, C., Liaw, A., Breiman, L., 2004. Using Random Forest to Learn Imbalanced Data. Technical Report (666), University of California, Berkeley, pp. 1–24.

Cohen-McFarlane, M., Goubran, R., Knoefel, F., 2020. Novel coronavirus cough database: NoCoCoDa. IEEE Access 8, 154087–154094.

Corman, V.M., Landt, O., Kaiser, M., Molenkamp, R., Meijer, A., Chu, D.K., Bleicker, T., Brünink, S., Schneider, J., Schmidt, M.L., et al., 2020. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Eurosurveillance 25.2000045 (3).

Cramer, J., Wu, H.-H., Salamon, J., Bello, J.P., 2019. Look, listen, and learn more: Design choices for deep audio embeddings. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3852–3856.

Das, R.K., Madhavi, M., Li, H., 2021. Diagnosis of COVID-19 using auditory acoustic cues. In: Proc. Interspeech 2021. pp. 921–925. http://dx.doi.org/10.21437/Interspeech.2021-497.

Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. 28 (4), 357–366.

Deshpande, G., Schuller, B.W., 2021. The DiCOVA 2021 challenge — An encoder-decoder approach for COVID-19 recognition from coughing audio. In: Proc. Interspeech 2021. pp. 931–935. http://dx.doi.org/10.21437/Interspeech.2021-811.

Elizalde, B., Tompkins, D., 2021. Covid-19 detection using recorded coughs in the 2021 dicova challenge. arXiv:2105.10619.

Eyben, F., Wöllmer, M., Schuller, B., 2010. Opensmile: The munich versatile and fast open-source audio feature extractor. In: Proc. 18th ACM Intl. Conf. Multimedia, pp. 1459–1462.

Harvill, J., Wani, Y.R., Hasegawa-Johnson, M., Ahuja, N., Beiser, D., Chestek, D., 2021. Classification of COVID-19 from cough using autoregressive predictive coding pretraining and spectral data augmentation. In: Proc. Interspeech 2021. pp. 926–930. http://dx.doi.org/10.21437/Interspeech.2021-799.

Hee, H.I., Balamurali, B., Karunakaran, A., Herremans, D., Teoh, O.H., Lee, K.P., Teng, S.S., Lui, S., Chen, J.M., 2019. Development of machine learning for asthmatic and healthy voluntary cough sounds: a proof of concept study. Appl. Sci. 9 (14), 2833.

Hershey, S., Chaudhuri, S., Ellis, D.P.W., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., Slaney, M., Weiss, R.J., Wilson, K., 2017. CNN Architectures for large-scale audio classification. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 131–135.

Imran, A., Posokhova, I., Qureshi, H.N., Masood, U., Riaz, M.S., Ali, K., John, C.N., Hussain, M.I., Nabeel, M., 2020. AI4Covid-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. Inform. Med. Unlocked 20, 100378.

Jaitly, N., Hinton, G.E., 2013. Vocal tract length perturbation (VTLP) improves speech recognition. In: International Conference on Machine Learning ICML, Vol. 117.

Kamble, M.R., Gonzalez-Lopez, J.A., Grau, T., Espin, J.M., Cascioli, L., Huang, Y., Gomez-Alanis, A., Patino, J., Font, R., Peinado, A.M., Gomez, A.M., Evans, N., Zuluaga, M.A., Todisco, M., 2021. PANACEA Cough sound-based diagnosis of COVID-19 for the DiCOVA 2021 challenge. In: Proc. Interspeech 2021. pp. 906–910. http://dx.doi.org/10.21437/Interspeech.2021-1062.

Kamble, M.R., Patil, H.A., 2019. Analysis of reverberation via teager energy features for replay spoof speech detection. In: In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 2607–2611.

Karas, V., Schuller, B.W., 2021. Recognising Covid-19 from coughing using ensembles of SVMs and lstms with handcrafted and deep audio features. In: Proc. Interspeech 2021. pp. 911–915. http://dx.doi.org/10.21437/Interspeech.2021-1267.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems, Vol. 30. Curran Associates, Inc..

Laguarta, J., Hueto, F., Subirana, B., 2020. COVID-19 artificial intelligence diagnosis using only cough recordings. IEEE Open J. Eng. Med. Biol. 1, 275–281.

Le, T.T., Fu, W., Moore, J.H., 2019. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. Bioinformatics 36 (1), 250–256.

Li, J., Huang, D.Q., Zou, B., Yang, H., Hui, W.Z., Rui, F., Yee, N.T.S., Liu, C., Nerurkar, S.N., Kai, J.C.Y., et al., 2021. Epidemiology of COVID-19: A systematic review and meta-analysis of clinical characteristics, risk factors, and outcomes. J. Med. Virol. 93 (3), 1449–1458.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2999–3007.

Mahanta, S.K., Jain, S., Kaushik, D., 2021. The Brogrammers DiCOVA 2021 Challenge System Report. Tech. Rep., DiCOVA Challenge, URL https://dicova2021.github.io/docs/reports/team_Brogrammers_DiCOVA_2021_Challenge_System_Report.pdf.

Mallol-Ragolta, A., Cuesta, H., Gómez, E., Schuller, B.W., 2021. Cough-based COVID-19 detection with contextual attention convolutional neural networks and gender information. In: Proc. Interspeech 2021. pp. 941–945. http://dx.doi.org/10.21437/Interspeech.2021-1052.

Muguli, A., Pinto, L., Nirmala, R., Sharma, N., Krishnan, P., Ghosh, P.K., Kumar, R., Bhat, S., Chetupalli, S.R., Ganapathy, S., Ramoji, S., Nanda, V., 2021. Dicova challenge: Dataset, task, and baseline system for COVID-19 diagnosis using acoustics. In: Proc. Interspeech 2021. pp. 901–905. http://dx.doi.org/10.21437/Interspeech.2021-74.

Oord, A.v.d., Li, Y., Vinyals, O., 2018. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

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), 1–10.

Park, D.S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E.D., Le, Q.V., 2019. Specaugment: A simple data augmentation method for automatic speech recognition. arXiv preprint arXiv:1904.08779.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.

Peeling, R.W., Olliaro, P.L., Boeras, D.I., Fongwen, N., 2021. Scaling up COVID-19 rapid antigen tests: promises and challenges. Lancet Infect. Dis..

Plakal, M., Ellis, D., 2020. Yamnet. https://github.com/tensorflow/models/tree/master/research/audioset/yamnet [Online; accessed on 15-June-2021].

Pramono, R.X.A., Imtiaz, S.A., Rodriguez-Villegas, E., 2016. A cough-based algorithm for automatic diagnosis of pertussis. PLoS One 11 (9), e0162128.

Ritwik, K.V.S., Kalluri, S.B., Vijayasenan, D., 2021. COVID-19 Detection from spectral features on the DiCOVA dataset. In: Proc. Interspeech 2021. pp. 936–940. http://dx.doi.org/10.21437/Interspeech.2021-1031.

Schaefer, I.-M., Padera, R.F., Solomon, I.H., Kanjilal, S., Hammer, M.M., Hornick, J.L., Sholl, L.M., 2020. In situ detection of SARS-CoV-2 in lungs and airways of patients with COVID-19. Mod. Pathol. 33 (11), 2104–2114.

Schuller, B.W., Batliner, A., Bergler, C., Messner, E.-M., Hamilton, A., Amiriparian, S., Baird, A., Rizos, G., Schmitt, M., Stappen, L., Baumeister, H., MacIntyre, A.D., Hantke, S., 2020. The interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks. In: Proc. Interspeech 2020. pp. 2042–2046. http://dx.doi.org/10.21437/Interspeech.2020-32.

Sharma, N., Krishnan, P., Kumar, R., Ramoji, S., Chetupalli, S.R., Nirmala, R., Ghosh, P.K., Ganapathy, S., 2020. Coswara – A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis. In: Proc. Interspeech, pp. 4811–4815.

Singh, V.P., Kumar, S., Jha, R.S., 2021. Samsung R&D Bangalore DiCOVA 2021 Challenge System Report. Tech. Rep., DiCOVA Challenge, URL https://dicova2021.github.io/docs/reports/team_samsung_DiCOVA_Technical_Report_IS2021.pdf.

Smith, J.O., Abel, J.S., 1999. Bark and ERB bilinear transforms. IEEE Trans. Speech Audio Process. 7 (6), 697–708.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S., 2018. X-Vectors: Robust DNN embeddings for speaker recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5329–5333.

Södergren, I., Nodeh, M.P., Chhipa, P.C., Nikolaidou, K., Kovács, G., 2021. Detecting COVID-19 from audio recording of coughs using random forests and support vector machines. In: Proc. Interspeech 2021. pp. 916–920. http://dx.doi.org/10.21437/Interspeech.2021-2191.

2020. Target product profiles for priority diagnostics to support response to the COVID-19 pandemic v.1.0 (WHO). https://www.who.int/docs/default-source/blueprint/who-rd-blueprint-diagnostics-tpp-final-v1-0-28-09-jc-ppc-final-cmp92616a80172344e4be0edf315b582021.pdf?sfvrsn=e3747f20_1&download=true [Online; accessed 20-May-2021].

2021. Virufy COVID-19 open cough dataset. https://github.com/virufy/virufy-data [Online; accessed 04-Jun-2021].

Xia, T., Han, J., Qendro, L., Dang, T., Mascolo, C., 2021. Uncertainty-aware COVID-19 detection from imbalanced sound data. In: Proc. Interspeech 2021. pp. 2951–2955. http://dx.doi.org/10.21437/Interspeech.2021-1320.

Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2017. Mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412.