THE SECOND DICOVA CHALLENGE: DATASET, TASK, AND BASELINE SYSTEM FOR COVID-19 DIAGNOSIS USING ACOUSTICS

Neeraj Kumar Sharma, Srikanth Raj Chetupalli, Debarpan Bhattachrya, Debottam Dutta, Pravin Mote, Sriram Ganapathy

email: sriramg@iisc.ac.in
Indian Institute of Science, Bangalore-560012, India

ABSTRACT

The Second DiCOVA Challenge aims at accelerating the research in diagnosing COVID-19 using acoustics (DiCOVA), a topic at the intersection of acoustics signal processing, machine learning, and healthcare. This challenge is an open call to researchers to analyze a dataset of audio recordings, collected from individuals with and without COVID-19, for a two-class classification. The development set audio recordings correspond to breathing, cough, and speech sound samples collected from 965 (172 COVID) individuals. The challenge features four tracks, one associated with each sound category and a fourth fusion track allowing experimentation with combination of the individual sound categories. In this paper, we introduce the challenge and provide a detailed description of the task and a baseline system.

Index Terms— COVID-19, acoustics, machine learning, respiratory diagnosis, healthcare

1. INTRODUCTION

Owing to the cost of machinery, expertise, and time, the molecular testing approaches, namely, the reverse transcription polymerase chain reaction (RT-PCR) test and the rapid antigen test (RAT) cannot be easily scaled up and deployed in a short time [1]. This results in a bottleneck in containing the spread of COVID-19, and has led to research on alternate testing methodologies [2]. Some of these include nanomaterial based bio-sensing of SARS-CoV-2 virus and radiographic imaging based on CT, X-ray, and ultrasound to categorize the health status of lungs. The primary symptoms of COVID-19 disease are fever, sore throat, cough, chest and muscle pain, and dyspnoea. Further, the pathogenesis of COVID-19 indicates minor to acute infection in the respiratory system during the onset of the disease. This has garnered interest in speech and audio research community, and there have been several studies [3, 4, 5, 6] gathering insights on the possibility of COVID-19 diagnosis using respiratory sounds. Acoustics based diagnosis can provide an excellent point-of-care, quick, easy to use, and cost-effective tool to detect COVID-19 infection, and consequently contain COVID-19 spread. Focusing on the progress made in acoustics based diagnosis of COVID-19, we think there is a need to benchmark the obtained detection performances.

The diagnosing COVID-19 using acoustics (DiCOVA) challenge series is designed to accelerate research along this direction. In this, a curated development dataset and baseline system is released, inviting researchers to build models surpassing the baseline on blind test set. A leaderboard style ranking is created for every evaluation on blind test set carried by the participants. The first DiCOVA Challenge[1] was a launched on 04-Feb-2021 and ran for 48 days. It garnered participation from 28 teams, coming from both academia and industry. Twenty one teams beat the baseline system performance. A summary of the challenge and systems developed by different teams is provided in [7]. Drawn by the interest of the research community in this topic lying at the interface of acoustic signal processing, machine learning, and healthcare, we have launched the **Second DiCOVA Challenge**[2] on 12-Aug-2021. This challenge features three distinct updates over the previous challenge. *Firstly*, since the closing of the first DiCOVA Challenge, there was a spike in daily global COVID-19 cases (Apr-May, 2021). The spike has been attributed to the new strains of the virus. This has helped us increase the data set size for the Second DiCOVA Challenge. *Secondly*, in addition to the cough sound category, the challenge brings to focus two additional sound categories, namely, breathing and speech. A leaderboard-style evaluation on blind test set is built for four tracks, one associated with each sound category (that is, breathing, cough, and speech) and a fourth fusion track allowing experimentation with combination of the individual sound categories. *Thirdly*, recently, multiple open datasets have been released to the public by different research groups. These include COVID-19 Sounds dataset [8] by University of Cambridge (UK), Buenos Aires COVID-19 Cough dataset [9] by Cabinet Ministers (Argentina), COUGHVID dataset [4] by EPFL University (Switzerland), and COVID-19 Open COUGH dataset [10] by Virufy (US). The participants are encouraged to use these datasets for enhancing model training and analysis.

Since its launch on 12−Aug-2021, the Second DiCOVA Challenge has received registration from 53 teams. These come from various countries (x% India, y %) and professional affiliations. In this paper, we present an overview of the dataset, tasks in the challenge, and the baseline system.

2. LITERATURE REVIEW

In a study by Imran et al. [3], a four class classifier is designed to detect healthy, pertussis, bronchitis, and COVID-19 individuals. On a privately collected cough sound dataset from hospitals, they report a sensitivity of 94% (and 91% specificity) using a convolutional neural network (CNN) architecture with mel-spectrogram feature input. Others studies have focused on the binary task of COVID-19 detection only. Agbley et al. [11] demonstrate 81% specificity (at 43% sensitivity) on a subset of the COUGHVID dataset [4]. Laguarte et al. [6] use a privately collected dataset of COVID-19 infected

[1] http://dicova2021.github.io/

[2] http://dicovachallenge.github.io/

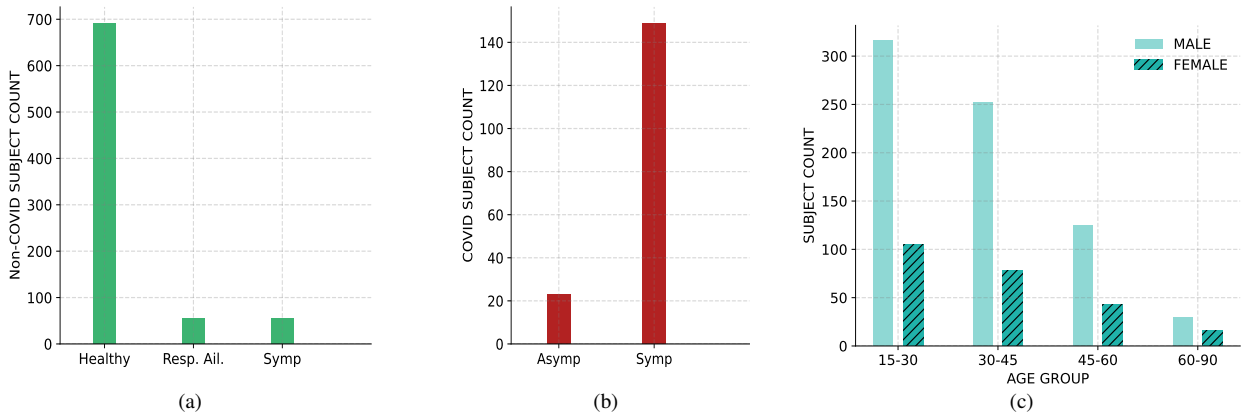


Fig. 1. Illustration of metadata corresponding to development dataset. (a) Health status of Non-COVID subjects broken down into categories of healthy (no symptoms), pre-existing respiratory ailment (asthma, chronic lung disease, pneumonia), and symptoms (cold, cough, fever, loss of taste or smell); (b) COVID-19 status COVID subjects; (c) Pooled subject gender and age distribution.

individuals and report an area under the-receiver operating characteristic curve (AUC-ROC) performance of 97.0%. Andreu-Perez et al. [12] create a controlled dataset by collecting cough sound samples from patients visiting hospitals, and report 98.8% AUC-ROC. A few studies have explored using the breathing and voice sounds as well. Brown et al. [5] created a dataset through crowd-sourcing, and analyzed COVID-19 detection. The authors report a performance between $80-82\%$ AUC-ROC. Han et al [13] propose using voice samples and demonstrate 77% AUC-ROC. Further, the use of symptoms along with voice provides a 2% improvement in AUC-ROC.
Altogether, these studies are encouraging. The limitations include: (i) a different COVID-19 patient population (with varied sizes) is used in each study, (ii) use of different performance evaluation methodologies, and (iii) lack of insight on acoustic feature differences between healthy and COVID-19 individuals. The Second DiCOVA Challenge is aimed at encouraging research groups to design and evaluate their classification system on a common dataset, using same performance metrics. We foresee this will facilitate obtaining benchmarks for system development in this topic.

3. DATASET

The challenge dataset is derived from the Coswara dataset [14], a crowd-sourced dataset of sound recordings collected by the authors. The Coswara data is collected using a website[3]. The volunteers from across the globe, age groups and health conditions were requested via social media campaigns to record their sound data in a quiet environment using an internet connected device (like, mobile phone or computer). Through the website, the subjects first provide demographic information like age and gender. This is followed by a short questionnaire to record their health status, including symptoms, pre-existing respiratory ailments, and co-morbidity, if any. Subsequently, their COVID-19 status is recorded via a questionnaire enquiring if they are currently COVID-19 positive, recovered, exposed to COVID-19 patients through primary contacts, or healthy. After collecting this information as metadata, the subjects records their acoustic data corresponding to 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 whole process takes $5-7$ minutes. The dataset collection protocol is approved by the Human Ethics Committee of the Indian Institute of Science, Bangalore (India).

The Second DiCOVA Challenge uses a subset of the Coswara dataset, sampled from the data collected between Apr-2020 and Jul-2021. The sampling included only age group of $15-90$ years. The subjects with health status of "recovered" (who were COVID positive and have recovered during the time of recording) and "exposed" (suspecting exposure to the virus) were not included in the dataset. Further, subjects with audio recordings of duration less than 500 msec were discarded. The resulting curated subject pool was divided into the following two groups. Only three sound categories are considered in the challenge. These correspond to breathing-deep, cough-heavy, and counting-normal, and are referred to as breathing, cough, and speech, respectively.

- **non-COVID**: Subjects self reported as healthy or having COVID-19 like symptoms (such as cold, cough, fever, muscle pain or fatigue, loss of taste or smell) or pre-existing respiratory ailments (such as 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 development dataset release is composed of audio records from 965 (172 COVID) subjects. This results in 965 (subjects)$\times$3 (sound categories) audio recordings. An illustration of the metadata of the subject pool is provided in Figure 1. About 75% of the subjects were male. The majority of the participants lie in the age group of $15-45$ years. In the non-COVID subject pool, close to 86% are healthy with no respiratory ailments and no COVID-19-like symptoms. The remaining 14% have respiratory ailments or COVID-19-like symptoms. In the COVID subject pool, close to 87% are symptomatic. In the development set, 17.2% subjects belong to COVID class. This represents an imbalanced dataset, reflecting the typical real-world scenario in the design of point-of-care-tests (POCTs) for COVID-19.
In the development data release shared with the participants, a five fold split corresponding to train and validation was also provided. This was done to facilitate hyper-parameter tuning using the same

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

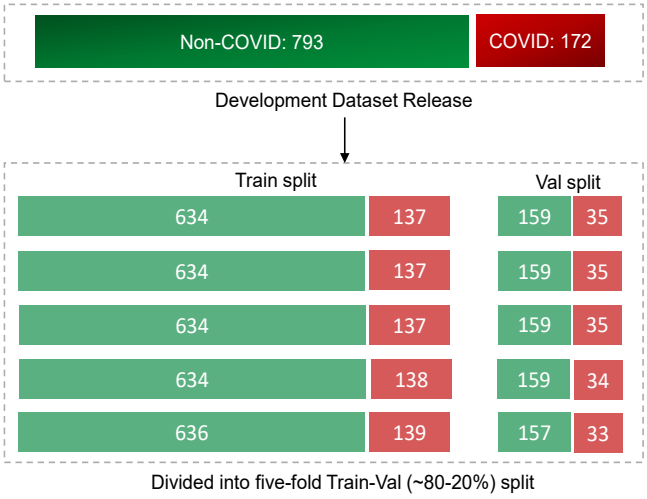


Fig. 2. Illustration of development dataset with depiction of train and val splits.

validation folds by all participants. The distribution COVID and non-COVID subjects in the splits is illustrated in Figure 2.

3.1. Audio Specifications

For the challenge, all audio recordings were re-sampled to 44.1 kHz and compressed into FLAC (Free Lossless Audio Codec) format for ease of distribution. The duration of audio in each track corresponded to 4.62 hrs for Track-1, 1.68 hrs for Track-2, and 3.93 hrs for Track-3, respectively.

4. CHALLENGE TASKS

The challenge task requires designing a binary classifier to detect COVID and Non-COVID subjects using sound categories corresponding to each track. Every registered team is provided with the development dataset to facilitate training and design of their own classification models. A team is free to use any dataset except the publicly available Coswara dataset[4] for data augmentation purposes. For evaluation, a blind test set comprising of 471 audio files $\times 3$ sound categories is provided the teams. An online website portal is created using Codalab[5]. The teams upload their probability scores on the blind test set audio files into this portal for evaluation. The portal computes the performance using the evaluation metrics and rank orders the team performance on a leaderboard in close to real-time. Every team gets maximum of 15 tickets for submitting scores to the leaderboard in each track.

Alongside the development dataset and blind test set, a baseline system is also shared with the participants. This provides a data processing pipeline tuned to the provided dataset. The participants can optionally design their system by modifying the baseline system. Post challenge, that is, 2-Oct-2021, all teams are required to submit their system report describing the designed models to the challenge organizers. Every team signed a terms and condition document[6] stating fair use of data and adhering to the rules of the challenge.

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

[5]https://competitions.codalab.org/competitions/34801

[6]https://dicovachallenge.github.io/docs/Second_DiCOVA_Challenge_Terms_Conditions_Doc.pdf

4.1. Evaluation Metrics

The focus of the challenge was on binary classification. As the dataset was imbalanced, we choose not to use accuracy as an evaluation metric. Each team submits COVID probability scores ($\in [0, 1]$, a higher value indicating a higher likelihood of COVID infection) for the list of validation and test audio recordings. We use the scores and the ground truth labels to compute the receiver operating characteristics (ROC) curve. The curve is obtained by varying the decision threshold between $0 - 1$ with a step size of 0.0001 and obtaining the specificity (true negative rate) and sensitivity (true positive rate) at every threshold value. The area under the resulting ROC curve, AUC-ROC, was used as a performance measure for the classifier. The area was computed using the trapezoidal method. The AUC-ROC is used as the primary evaluation metric. Further, specificity at a sensitivity greater than or equal to 95% was used as a secondary evaluation metric. For brevity, we will refer to AUC-ROC by AUC in the rest of the paper.

5. BASELINE SYSTEM

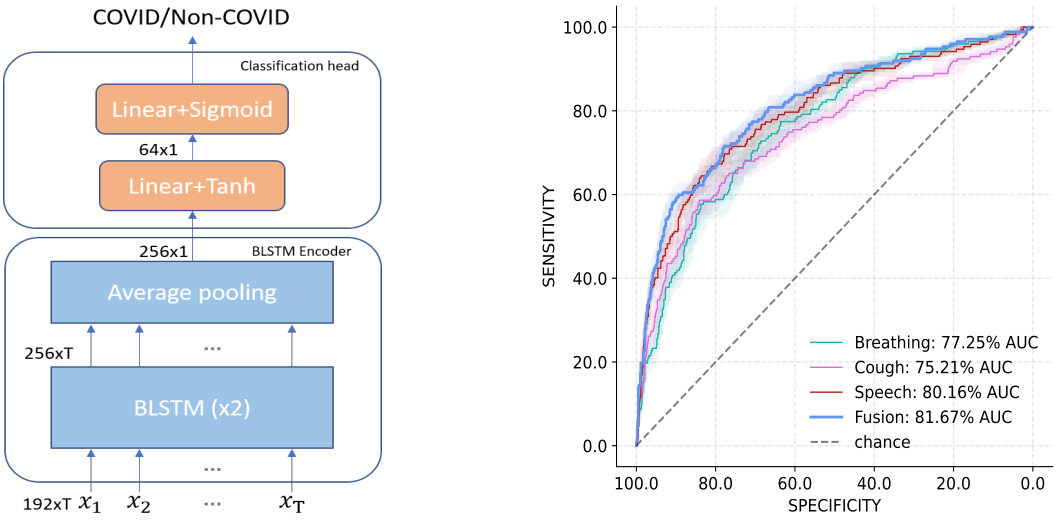
The baseline system is composed of the following processing stages. The stages are implemented using python and make use of the scikit-learn [15] and the PyTorch python packages [16]. The same baseline system is used for all tracks.

Pre-processing: First the audio sample inside the sound file are normalized to lie between $\pm$ 1. This is followed by discarding low activity regions from the signal. Using a sound sample activity detection threshold of 0.01 and a buffer size of 50 msec on either size of a sample, any audio sample with less than the threshold is discarded.

Feature extraction: The log mel-spectrogram features are extracted using short-time windowed segments of size 1024 samples (23.2 msec) and temporal hop of 441 samples (10 ms), and a 64 mel-filter bank. This results in a $64 \times N_k$ dimensional feature matrix for the k^{th} sound file, where N_k represents the number of short-time frames. The mel-spectrogram features are appended with the first and second order temporal derivatives. The resulting $192 \times N_k$ dimensional features are file-level mean-variance normalized.

Classifier: Initial experimentation with traditional classifiers such as logistic regression and random forest gave an AUC performance in the range $60 - 70\%$ AUC on the validation folds for all the tracks. With an aim to provide a better baseline performance, we opted to using deep neural network architecture. A cascade of two bi-directional long-short term memory (LSTM) and a fully connected neural network was used. This is an encoder-decoder style network and an illustration is provided in Figure 3(a). The encoder consists of two BLSTM layers with 128 units in both the forward and backward direction. The $256 \times T$ dimensional output of the BLSTM layers is fed to a pooling layer which performs averaging along the time dimension to generate a sequence level embedding of 256×1. This output is interfaced with decoder. This is fully connected neural network comprising of 256 nodes in the first layer and 64 nodes and a $\tanh(\cdot)$ non-linearity in the second layer. Finally, a single node output, passed through a sigmoid non-linearity is obtained as the COVID probability score for the input feature matrix.

Training: We extract contiguous T frame chunks, with a 10 frame hop, from the features matrices to obtain $192 \times T$ fixed dimensional representations for training the classifier. We choose T as 51. The label of each chunk is same as that of the audio file. Each mini-batch is composed of 1024 numbers of $192 \times T$ chunks, randomly sampled from different audio files such that the proportion of



(a) (b)

Fig. 3. Illustration of: (a) Baseline system architecture, and (b) average ROC obtained on the validation folds.

Validation	AUC-ROC Performance (in %)			
	Breathing	**Cough**	**Speech**	**Fusion**
fold-0	74.82	71.77	75.35	77.25
fold-1	73.94	78.18	87.16	82.35
fold-2	74.26	77.21	80.59	81.83
fold-3	79.99	74.03	78.22	80.29
fold-4	83.25	74.87	79.46	86.63
Average	**77.25**	**75.21**	**80.16**	**81.67**

Table 1. Baseline System performance on the validation folds provided in the development dataset.

Test	AUC-ROC Performance (in %)			
	Breathing	**Cough**	**Speech**	**Fusion**
Blind	84.50	74.89	84.26	84.70

Table 2. Baseline System performance on the blind test set.

COVID and non-COVID labels is close to 0.5. This takes care of class imbalances by oversampling the minority (COVID) class. The binary cross entropy (BCE) loss, and Adam optimizer with an initial learning rate of 0.0001 and ℓ_2 regularization set to 0.0001 is used to train the classifier for 10 epochs. The learning rate is reduced by a factor of 10 if the validation loss is increasing for three consecutive epochs. A dropout factor of 0.1 is applied to the outputs of first BLSTM layer and the first linear layer.

Inference: Given an audio recording, $192 \times T$ mel-spectrogram feature chunks (with hop of 10 frames) are extracted (similar to the training stage). These are input to the trained classifier and the output probability scores are obtained for each chunk. The average of the chunk probability scores is output as the COVID probability score of the audio file.

Fusion: Three classifiers are trained separately for the three track-wise sound categories, namely, breathing, cough and speech, respectively. A final prediction at the subject level is obtained by averaging the COVID probability scores for the three audio files corresponding to the subject.

6. RESULTS

Here we present the baseline system performance. The average ROC, computed over the five validation folds, for each sound category are shown in Figure 3(b). The AUC performance is significantly better than chance (50% AUC) for all sound categories. This suggests that there is information in the acoustic features to help discriminating COVID from Non-COVID subjects. Table 1 depicts the AUCs obtained on each validation fold. For each fold (shown in Fig. 2), the classifier is trained using the data from train split and evaluated on the data from validation split. On three out of five folds, speech sound category gave a better AUC than breathing and cough. A fusion of scores from all three sound categories resulted in an improvement in four out of five folds. An average AUC performance of 77.25%, 75.21% and ,80.16% was obtained for breathing, cough, and speech categories, respectively. The best average AUC of 81.67%was obtained for the Fusion track.

On the blind test set, for each sound category we draw inference by averaging the score obtained using model trained on each train fold. Amongst the three sound categories, the best test set AUC, equating to 84.50%, was obtained for breathing sound category. This was followed by speech and cough. A fusion resulted in a slight improvement with an AUC 84.70%.

7. CONCLUSION

The uniqueness of the dataset makes the Second DiCOVA Challenge a first-of-its kind in the speech and audio research community. The practical and timely relevance of the task encourages a focused ef-

fort from researchers across the globe, and from diverse fields such as respiratory sciences, speech and audio processing, and machine learning. Along with the dataset, we also provide the baseline system software to all the participants. We expect this will serve as an example data processing pipeline for the participants. Further, participants are encouraged to explore different kinds of features and models of their own choice to obtain significantly better performance compared to the baseline system.

8. ACKNOWLEDGEMENT

The authors would like to thank Anand Mohan for his enormous help in website design and data collection efforts, and the Department of Science and Technology, Government of India for funding the Coswara Project through the RAKSHAK Programme.

9. REFERENCES

[1] Tim R Mercer and Marc Salit, "Testing at scale during the covid-19 pandemic," *Nature Reviews Genetics*, vol. 22, no. 7, pp. 415–426, 2021.

[2] Bhavesh D Kevadiya, Jatin Machhi, Jonathan Herskovitz, Maxim D Oleynikov, Wilson R Blomberg, Neha Bajwa, Dhruvkumar Soni, Srijanee Das, Mahmudul Hasan, Milankumar Patel, et al., "Diagnostics for SARS-CoV-2 infections," *Nature materials*, pp. 1–13, 2021.

[3] Ali Imran, Iryna Posokhova, Haneya N. Qureshi, Usama Masood, Muhammad Sajid Riaz, Kamran Ali, Charles N. John, MD Iftikhar Hussain, and Muhammad Nabeel, "AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app," *Informatics in Medicine Unlocked*, vol. 20, pp. 100378, 2020.

[4] Lara Orlandic, Tomas Teijeiro, and David Atienza, "The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms," *Scientific Data*, vol. 8, no. 1, pp. 1–10, 2021.

[5] Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Jing Han, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, and Cecilia Mascolo, "Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data," in *Proc. 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, New York, NY, USA, 2020, p. 3474–3484, Association for Computing Machinery.

[6] Jordi Laguarta, Ferran Hueto, and Brian Subirana, "COVID-19 artificial intelligence diagnosis using only cough recordings," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 1, pp. 275–281, 2020.

[7] Neeraj Kumar Sharma, Ananya Muguli, Prashant Krishnan, Rohit Kumar, Srikanth Raj Chetupalli, and Sriram Ganapathy, "Towards sound based testing of covid-19–summary of the first diagnostics of covid-19 using acoustics (dicova) challenge," *arXiv preprint arXiv:2106.10997*, 2021.

[8] "Cambridge University, UK - COVID-19 Sounds App," https://www.covid-19-sounds.org/en/blog/data_sharing.html, 2020, [Online; accessed 16-Aug-2021].

[9] "Buenos Aires COVID-19 Cough Data Dataset," https://data.buenosaires.gob.ar/dataset/tos-covid-19/, 2021, [Online; accessed 16-Aug-2021].

[10] "Virufy COVID-19 Open Cough Dataset," https://github.com/virufy/virufy-data, 2021, [Online; accessed 04-Jun-2021].

[11] Bless Lord Y Agbley, Jianping Li, Aminul Haq, Bernard Cobbinah, Delanyo Kulevome, Priscilla A Agbefu, and Bright Eleeza, "Wavelet-based cough signal decomposition for multimodal classification," in *17th Intl. Computer Conference on Wavelet Active Media Technology and Information Processing)*. IEEE, 2020, pp. 5–9.

[12] Javier Andreu-Perez, Humberto Perez-Espinosa, Eva Timonet, Mehrin Kiani, Manuel Ivan Giron-Perez, Alma B. Benitez-Trinidad, Delaram Jarchi, Alejandro Rosales, Nick Gkatzoulis, Orion F. Reyes-Galaviz, Alejandro Torres, Carlos Alberto Reyes-Garcia, Zulfiqar Ali, and Francisco Rivas, "A generic deep learning based cough analysis system from clinically validated samples for point-of-need COVID-19 test and severity levels," *IEEE Trans. Services Computing*, pp. 1–1, 2021.

[13] Jing Han, Chloë Brown, Jagmohan Chauhan, Andreas Grammenos, Apinan Hasthanasombat, Dimitris Spathis, Tong Xia, Pietro Cicuta, and Cecilia Mascolo, "Exploring automatic covid-19 diagnosis via voice and symptoms from crowdsourced data," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 8328–8332.

[14] Neeraj Sharma, Prashant Krishnan, Rohit Kumar, Shreyas Ramoji, Srikanth Raj Chetupalli, R Nirmala, Prasanta Kumar Ghosh, and Sriram Ganapathy, "Coswara – a database of breathing, cough, and voice sounds for COVID-19 diagnosis," in *Proc. Interspeech*, 2020, pp. 4811–4815.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[16] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., pp. 8024–8035. Curran Associates, Inc., 2019.