Evaluating the COVID-19 Identification ResNet (CIdeR) on the INTERSPEECH COVID-19 From Audio Challenges

Alican Akman[1][*][†], Harry Coppock[1][†], Alexander Gaskell[1], Panagiotis Tzirakis[1], Lyn Jones[2] and Björn W. Schuller[1,3]

[1] GLAM–Group on Language, Audio, and Music, Imperial College London, London, United Kingdom, [2] Department of Radiology, North Bristol NHS Trust, Bristol, United Kingdom, [3] Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Augsburg, Germany

Several machine learning-based COVID-19 classifiers exploiting vocal biomarkers of COVID-19 has been proposed recently as digital mass testing methods. Although these classifiers have shown strong performances on the datasets on which they are trained, their methodological adaptation to new datasets with different modalities has not been explored. We report on cross-running the modified version of recent COVID-19 Identification ResNet (CIdeR) on the two Interspeech 2021 COVID-19 diagnosis from cough and speech audio challenges: ComParE and DiCOVA. CIdeR is an end-to-end deep learning neural network originally designed to classify whether an individual is COVID-19-positive or COVID-19-negative based on coughing and breathing audio recordings from a published crowdsourced dataset. In the current study, we demonstrate the potential of CIdeR at binary COVID-19 diagnosis from both the COVID-19 Cough and Speech Sub-Challenges of INTERSPEECH 2021, ComParE and DiCOVA. CIdeR achieves significant improvements over several baselines. We also present the results of the cross dataset experiments with CIdeR that show the limitations of using the current COVID-19 datasets jointly to build a collective COVID-19 classifier.

Keywords: COVID-19, computer audition, digital health, deep learning, audio

1. INTRODUCTION

The current coronavirus pandemic (COVID-19), caused by the severe-acute-respiratory-syndrome-coronavirus 2 (SARS-CoV-2), has infected a confirmed 126 million people and resulted in 2,776,175 deaths (WHO)[1]. Mass testing schemes offer the option to monitor and implement a selective isolation policy to control the pandemic without the need for regional or national lockdown (1). However, physical mass testing methods, such as the Lateral Flow Test (LFT) have come under criticism since the tests divert limited resources from more critical services (2, 3) and due to suboptimal diagnostic accuracy. Sensitivities of 58 % have been reported for self-administered LFTs (4), unacceptably low when used to detect active virus, a context where high sensitivity is essential to prevent the reintegration into society

---

[1] As of 29th March 2021 https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

of falsely reassured infected test recipients (5). In addition to mass testing, radar remote life sensing technology offers non-contact applications to combat COVID-19 including heart rate tracking, identity authentication, indoor monitoring and gesture recognition (6).

Investigating the potential for digital mass testing methods is an alternative approach, based on findings that suggest a biological basis for identifiable vocal biomarkers caused by SARS-CoV-2's effects on the lower respiratory track (7). This has recently been backed up by empirical evidence (8). Efforts have been made to collect and classify a range of different modality audio recordings of COVID-19-positive and COVID-19-negative individuals and several datasets have been released that use applications to collect the breath and cough of volunteer individuals. Examples include the "COUGHVID" (9), "Breath for Science"[2], "Coswara" (10), COVID-19 sounds[3], and 'CoughAgainstCovid' (11). In addition, to focus the attention of the audio processing community onto the task of binary classification of COVID-19 from audio, two INTERSPEECH competitions: the INTERSPEECH 2021 Computational Paralinguists Challenge (ComParE)[4] (12) with its COVID-19 Cough and Speech Sub-Challenges, and Diagnosing COVID-19 using acoustics (DiCOVA)[5] (13) have been organized with this focus as their challenge.

Several studies have been published that propose machine learning-based COVID-19 classifiers exploiting distinctive sound properties between positive and negative cases to classify these datasets. Brown et al. (14) and Ritwik et al. (15) demonstrate that simple machine learning models perform well in these relatively small datasets. In addition, deep neural networks are exploited in Laguarta et al. (16), Pinkas et al. (17), Imran et al. (18), and Nessiem et al. (19) with proven performance at the COVID-19 detection task. Although there are works that try to combine different modalities computing the representations separately, Coppock et al. (20) (CIdeR) proposes an approach computing joint representation of a number of modalities. The adaptability of this approach to different types of datasets has not to our knowledge been explored or reported.

To this end, we propose a modified version of COVID-19 Identification ResNet (CIdeR), a recently developed end-to-end deep learning neural network optimized for binary COVID-19 diagnosis from cough and breath audio (20), which is applicable to common datasets with further modalities. We present the competitive results of CIdeR to the two COVID-19 cough and speech Challenges of INTERSPEECH 2021, ComParE and DiCOVA. We also investigate the behavior of a strong COVID-19 classifier across different datasets by running cross dataset experiments with CIdeR. We describe the limitations of current COVID-19 classifiers with these experiments regarding the ultimate goal of proposing a universal COVID-19 classifier.

2. METHODS

2.1. Model

CIdeR (20) is a 9 layer convolutional residual network. A schematic detailing of the model can be seen in **Figure 1**. Each layer or block consists of a stack of convolutional layers with Rectified Linear Units (ReLUs). Batch normalization (21) also features in the residual units, acting as a source of regularization and supporting training stability. A fully connected layer with sigmoid activation terminates the model yielding a single logit output which can be interpreted as an estimation of the probability of COVID-19. As detailed in **Figure 1** the network is modified to be compatible with a varying number of modalities, for example, if a participant has provided cough, deep breathing, and sustained vowel phonation audio recordings, they can be stacked in a depth wise manner and passed through the network as a single instance. We use PyTorch library in python to implement CIdeR and baseline models.
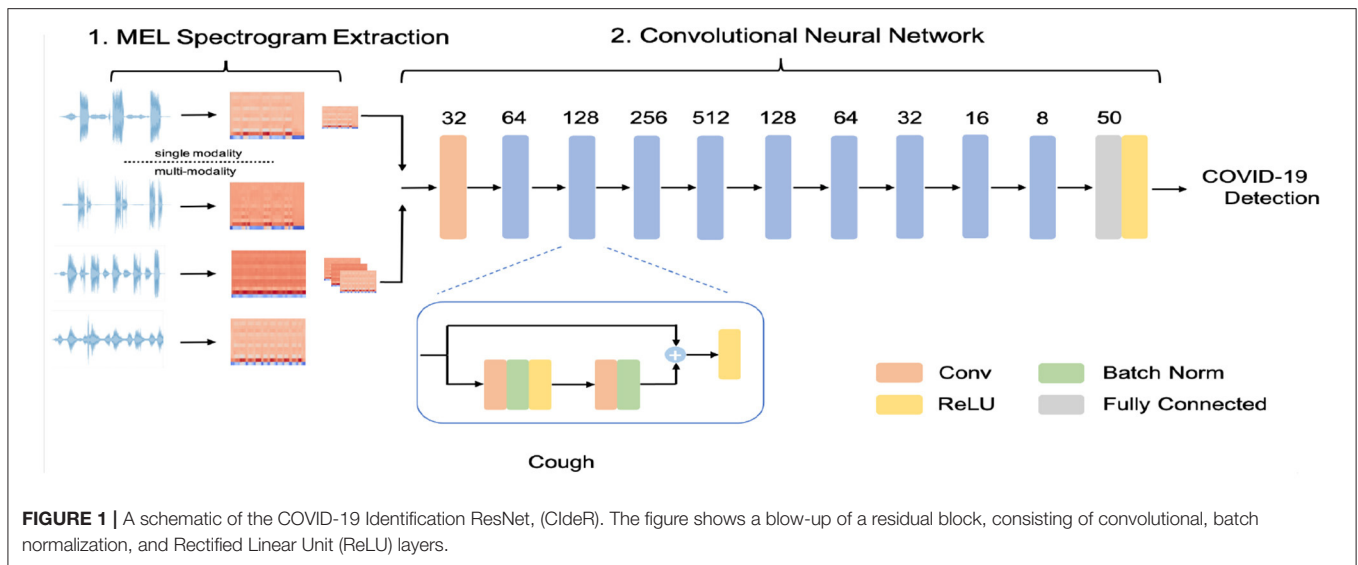
2.2. Pre-processing

At training time, a window of s-seconds, which was fixed at 6 s for these challenges, is sampled from the audio recording randomly. If the audio recording is less than s-seconds long, the sample is padded with repeated versions of itself. The sampled audio is then converted into Mel-Frequency Cepstral Coefficients (MFCCs) resulting in an image of width s * the sample rate and height equal to the number of MFCCs. Three data augmentation steps are then applied to the sample. First, the pitch of the recording is randomly shifted, secondly, bands of the Mel spectrogram are masked in the time and Mel coefficient axes and finally, Gaussian noise is added. At test time, the sampled audio recording is chunked into a set of s-second clips and processed in parallel. The mean of the set of logits is then returned as the final prediction.

2.3. Baselines

The DiCOVA team ran baseline experiments for the track 1 (coughing) sub-challenge; only the best performing (MLP) model's score was reported. For the track 2 (deep breathing/vowel phonation/counting) sub-challenge, however, baseline results were not provided. Baseline results were provided for the ComParE challenge but only Unweighted Average Recall (UAR) was reported rather than Area Under Curve of the Receiver Operating Characteristics curve (ROC-(AUC)). To allow comparison across challenges, we created new baseline results for the ComParE sub-challenges and the DiCOVA Track 2 sub-challenge, using the same baseline methods described for the DiCOVA Track 1 sub-challenge. The three baseline models applied to all four sub-challenge datasets were Logistic Regression (LR), Multi-layer Perceptron (MLP), and Random Forrest (RF), where the same hyperparameter configurations that were specified in the DiCOVA baseline algorithm was used (13).

To provide a baseline comparison for the CIdeR track 2 results, we built a multimodal baseline model. We followed a similar strategy with the provided DiCOVA baseline algorithm, while extracting the features for each modality. Rather than individual training for different models, we developed an algorithm that concatenates input features from separate

FIGURE 1 | A schematic of the COVID-19 Identification ResNet, (CIdeR). The figure shows a blow-up of a residual block, consisting of convolutional, batch normalization, and Rectified Linear Unit (ReLU) layers.

TABLE 1 | ComParE sub-challenge dataset splits.

#	CCS			CSS		
	Train	**Val**	**Test**	**Train**	**Val**	**Test**
COVID-19-postive	71	48	39	72	142	94
COVID-19-negative	215	183	169	243	153	189
Total	286	231	208	315	295	283

Values specify the number of audio recordings, not the number of participants.

TABLE 2 | DiCOVA sub-challenge dataset splits.

#	Track-1		Track-2	
	Train + Val	**Test**	**Train + val**	**Test**
COVID-19-postive	75	blind	60	21
COVID-19-negative	965	blind	930	188
Total	1,040	234	990	209

The test set labels were withheld by the DiCOVA team, contestants had to submit predictions for each test case, on which a final AUC was returned.

modalities. Then, this combined feature set was fed to the baseline models: LR, MLP, and RF.

We used 39 dimensional MFCCs as our feature type to represent the input sounds. For LR, we used Least Square Error (L2) as a penalty term. For MLP, we used a single hidden layer of size 25 with a Tanh activation layer and L2 regularization. The Adam optimiser and a learning rate of 0.0001 was used. For RF, we built the model with 50 trees and split based on the gini impurity criterion.

3. DATASETS

3.1. ComParE

ComParE hosted two COVID-19 related sub-challenges, the COVID-19 Cough Sub-Challenge (CCS) and the COVID-19 Speech Sub-Challenge (CSS). Both CCS and CSS are subsets of the crowd sourced Cambridge COVID-19 sound database (14, 22). CCS consists of 926 cough recordings from 397 participants. Participants provided 1–3 forced coughs resulting in a total of 1.63 h of recording. CSS is made up of 893 recordings from 366 participants totalling 3.24 h of recording. Participants were asked to recite the phrase *"I hope my data can help manage the virus pandemic"* in their native language 1–3 times. The train-test splits for both sub-challenges are detailed in **Table 1**.

3.2. DiCOVA

Once again, DiCOVA hosted two COVID-19 audio diagnostic sub-challenges. Both sub-challenge datasets were subsets of the crowd sourced Coswara dataset (10). The first sub-challenge, named Track-1, comprised of a set of 1,274 forced cough audio recordings from 1,274 individuals totalling 1.66 h. The second, Track-2, was a multi-modality challenge, where 1,199 individuals provided three separate audio recordings; deep breathing, sustained vowel phonation, and counting from 1 to 20. This dataset represented a total of 14.9 h of recording. The train-test splits are detailed in **Table 2**.

4. RESULTS AND DISCUSSION

The results from the array of experiments with CIdeR and the 3 baseline models are detailed in **Table 3**. CIdeR performed strongly across all four sub-challenges, achieving AUCs of 0.799 and 0.787 in the DiCOVA Track 1 and 2 sub-challenges, respectively, and 0.732 and 0.787 in the ComParE CCS and CSS sub-challenges. In the DiCOVA cough sub-challenge, CIdeR significantly outperformed all three baseline models based on 95 % confidence intervals calculated following (23), and in the DiCOVA breathing and speech sub-challenge it achieved a

TABLE 3 | Results for CIdeR and a range of baseline models for 4 sub-challenges across the DiCOVA and ComParE challenges.

	Sub-challenge*	CIdeR	Baseline		
			LR	MLP	RF
DiCOVA	Track 1**	**0.799** ± 0.058	-	0.699 ± 0.068	-
	Track 2	0.786 ± 0.057	0.647 ± 0.014	0.684 ± 0.072	0.776 ± 0.063
ComParE	CCS	0.732 ± 0.068	0.722 ± 0.069	0.765 ± 0.065	0.753 ± 0.066
	CSS	**0.787** ± 0.060	0.583 ± 0.072	0.656 ± 0.070	0.628 ± 0.070

Testing is performed on the held-out test fold once final model decisions have been made on the validation sets. The Area Under Curve of the Receiver Operating Characteristics curve (AUC(-ROC)) is displayed. A 95% confidence interval is also shown following (23). CIdeR scores which are statistically higher than the best baseline results with a 95% confidence are in bold. The three baseline models are Logistic Regression (LR), Multi-layer Perceptron (MLP), and Random Forrest (RF). All baseline models were trained on MFCC features.
*Track 1: coughing, Track 2: deep breathing + vowel phonation + counting, CCS: coughing, CSS: speech—" hope my data can help managethe virus pandemic". ** As the demographics were not provided for the Track 1 test set, when calculating the AUC confidence intervals, it was assumed that there was an equal number of COVID-19-positive and COVID-19-negative recordings.

higher AUC although the improvement over the baselines was not significant. Conversely, while CIdeR performed significantly better than all three baseline models in the ComParE speech sub-challenge based on 95 % confidence intervals calculated following (23), it performed no better than baseline in the ComParE cough sub-challenge. One can speculate that this may have resulted from the small dataset sizes favoring the more classical machine learning approaches which do not need as much training data.

4.1. Limitations

A key limitation with both the ComParE and DiCOVA COVID-19 challenges is the size of the datasets. Both datasets contain very few COVID-19-positive participants. Therefore, the certainty in results is limited and this is reflected in the large 95 % confidence intervals detailed in **Table 3**. This issue is compounded by the demographics of the datasets. As detailed in Brown et al. (14) and Muguli et al. (13) for the ComParE datasets and the DiCOVA datasets, respectively, not all demographics from society are represented evenly—most notably, there is poor coverage of age and ethnicity and both datasets are skewed toward the male gender.

In addition, the crowd-sourced nature of the datasets introduces some confounding variables. Audio is a tricky sense to control. It contains a lot of information about the surrounding environment. As both datasets were crowd-sourced, there could have been correlations between ambient sounds and COVID-19 status, for example, sounds characteristic of hospitals or intensive care units being more often present for COVID-19-positive recordings compared to COVID-19-negative recordings. As the ground truth labels for both datasets were self reported, presumably the participants knew at the time of recording whether they had COVID-19 or not. One could postulate that the individuals who knew they were COVID-19-positive might have been more fearful than COVID-19-negative participants at the time of recording, an audio characteristic known to be identifiable by machine learning models (24). Therefore, the audio features which have been identified by the model may not be specific audio biomarkers for the disease.

We note that both the DiCOVA Track 1 and ComParE CCS sub-challenges were cough recordings. Therefore, there was an opportunity to utilize both training sets. Despite having access

TABLE 4 | The results for cross dataset experiments.

	Test set		
Train set	DiCOVA	ComParE	COUGHVID
DiCOVA	0.799	0.554	0.464
ComParE	0.512	0.732	0.552
COUGHVID	0.395	0.518	0.566
All	0.673	0.717	0.531

to both the DiCOVA and ComParE datasets, training on the two datasets together did not yield a better performance on either of the challenges' test sets. Additionally, a model which performed well on one of the challenges test sets would see a marked drop in performance on the other challenge's test set. We run cross dataset experiments to analyse this effect further. For these experiments, we also included the COUGHVID dataset (9) in which COVID-19 labels were assigned by experts and not as a results of clinically validated test. The results in **Table 4** show that the trained models for each dataset do not generalize well and perform poorly on excluded datasets. This is a worrying find, as it suggests that audio markers which are useful in COVID-19 classification in one dataset are not useful or present in the other dataset. This agrees with the concerns presented in Coppock et al. (25) that current COVID-19 audio datasets are plagued with bias, allowing for machine learning models to infer COVID-19 status, not by audio biomarkers uniquely produced by COVID-19, but by other correlations in the dataset such as nationality, comorbidity and background noise.

5. FUTURE WORK

One of the most important next steps is to collect and evaluate machine learning COVID-19 classification on a larger dataset that is more representative of the population. To achieve optimal ground truth, audio recordings should be collected at the time that the Polymerase Chain Reaction (PCR) test is taken, before the result is known. This would ensure full blinding of the participant to their COVID-19 status and exclude any

environmental audio biasing in the dataset. The Cycle Threshold (CT) of the PCR test should also be recorded, CT correlates with viral load [26] and therefore would enable researchers to determine the model's classification performance to the disease at varying viral loads. This relationship is critical in assessing the usefulness of any model in the context of a mass testing scheme, since the ideal model would detect a viral load lower than the level that confers infectiousness[6]. Finally, studies similar to Bartl-Pokorny et al. [8], directly comparing acoustic features of COVID-19-positive and COVID-19-negative participants should be conducted on all publicly available datasets.

6. CONCLUSION

Cross-running CIdeR on the two 2021 Interspeech COVID-19 diagnosis from cough and speech audio challenges has demonstrated the model's adaptability across multiple modalities. With little modification, CIdeR achieves competitive results in all challenges, advocating the use of end-2-end deep learning models for audio processing thanks to their flexibility and strong performance. Cross dataset experiments with CIdeR have revealed the technical limitations of the datasets for joint usage that prevent from creating a common COVID-19 classifier.

---

[6]Seventy-third SAGE meeting on COVID-19, 17th December 2020.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

AA and HC designing alternative methods, literature analysis, performing and analyzing experiments, manuscript preparation, editing, and drafting manuscript. AG and PT designing and performing experiments. BS and LJ drafting manuscript and manuscript editing. All authors revised, developed, read, and approved the final manuscript.

FUNDING

REFERENCES

1. Peto J, Carpenter J, Smith GD, Duffy S, Houlston R, Hunter DJ, et al. Weekly COVID-19 testing with household quarantine and contact tracing is feasible and would probably end the epidemic. *R Soc Open sci*. (2020) 8:201546. doi: 10.1098/rsos.200915
2. Wise J. Covid-19: concerns persist about purpose, ethics, and effect of rapid testing in Liverpool. *BMJ*. (2020) 371:m4690. doi: 10.1136/bmj.m4690
3. Holt E. Newsdesk COVID-19 testing in Slovakia. *Lancet Infect Dis*. (2021) 21:32. doi: 10.1016/S1473-3099(20)30948-8
4. Mahase E. Covid-19: innova lateral flow test is not fit for test and release strategy, say experts. *BMJ*. (2020) 371:m4469. doi: 10.1136/bmj.m4469
5. Moyse KA. Lateral flow tests need low false positives for antibodies and low false negatives for virus. *BMJ*. (2021) 372:n90. doi: 10.1136/bmj.n90
6. Islam SMM, Fioranelli F, Lubecke VM. Can radar remote life sensing technology help combat COVID-19? *Front Commun Netw*. (2021) 2:648181. doi: 10.3389/frcmn.2021.648181
7. Quatieri T, Talkar T, Palmer J. A framework for biomarkers of COVID-19 based on coordination of speech-production subsystems. *IEEE Open J Eng Med Biol*. (2020) 1:203–6. doi: 10.1109/OJEMB.2020.2998051
8. Bartl-Pokorny KD, Pokorny FB, Batliner A, Amiriparian S, Semertzidou A, Eyben F, et al. The voice of COVID-19: acoustic correlates of infection in sustained vowels. *J Acoust Soc Am*. (2020) 149:4377. doi: 10.1121/10.0005194
9. Orlandic L, Teijeiro T, Atienza D. The COUGHVID crowdsourcing dataset: a corpus for the study of large-scale cough analysis algorithms. *Sci Data*. (2020) 8:156. doi: 10.1038/s41597-021-00937-4
10. Sharma N, Krishnan P, Kumar R, Ramoji S, Chetupalli SR, R N, et al. Coswara-a database of breathing, cough, and voice sounds for COVID-19 diagnosis. In: *Proceedings Interspeech*. Shanghai (2020). p. 4811–5.
11. Bagad P, Dalmia A, Doshi J, Nagrani A, Bhamare P, Mahale A, et al. Cough against COVID: evidence of COVID-19 signature in cough sounds. *arXiv*. (2020). doi: 10.48550/arXiv.2009.08790
12. Schuller BW, Batliner A, Bergler C, Mascolo C, Han J, Lefter I, et al. The INTERSPEECH 2021 computational paralinguistics challenge: COVID-19

cough, COVID-19 speech, escalation & primates. *arXiv:2102.13468*. (2021) doi: 10.21437/Interspeech.2021-19
13. Muguli A, Pinto L, Nirmala R, Sharma N, Krishnan P, Ghosh PK, et al. DiCOVA challenge: Dataset, task, and baseline system for COVID-19 diagnosis using acoustics. *arXiv [Preprint]*. (2021). arXiv: 2103.09148. doi: 10.48550/ARXIV.2103.09148
14. Brown C, Chauhan J, Grammenos A, Han J, Hasthanasombat A, Spathis D, et al. Exploring automatic diagnosis of COVID-19 from crowdsourced respiratory sound data. In: *Proceedings of Knowledge Discovery and Data Mining*. (2020). p. 3474–84.
15. Ritwik KVS, Kalluri SB, Vijayasenan D. COVID-19 patient detection from telephone quality speech data. *arXiv*. (2020). doi: 10.48550/arXiv.2011.04299
16. Laguarta J, Hueto F, Subirana B. COVID-19 artificial intelligence diagnosis using only cough recordings. *IEEE Open J Eng Med Biol*. (2020) 1:275–81. doi: 10.1109/OJEMB.2020.3026928
17. Pinkas G, Karny Y, Malachi A, Barkai G, Bachar G, Aharonson V. SARS-CoV-2 detection from voice. *IEEE Open J Eng Med Biol*. (2020) 1:268–74. doi: 10.1109/OJEMB.2020.3026468
18. Imran A, Posokhova I, Qureshi HN, Masood U, Riaz S, Ali K, et al. AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. *arXiv*. (2020). doi: 10.1016/j.imu.2020.100378
19. Nessiem MA, Mohamed MM, Coppock H, Gaskell A, Schuller BW. Detecting COVID-19 from breathing and coughing sounds using deep neural networks. In: *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*. Aveiro: IEEE (2021). p. 183–8.
20. Coppock H, Gaskell A, Tzirakis P, Baird A, Jones L, Schuller BW. End-2-End COVID-19 detection from breath & cough audio. *BMJ Innovations*. (2021). doi: 10.1136/bmjinnov-2021-000668
21. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: *Proceedings of International Conference on Machine Learning. vol. 37*. Lille (2015). p. 448–56.
22. Han J, Brown C, Chauhan J, Grammenos A, Hasthanasombat A, Spathis D, et al. 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)*. Toronto, ON: IEEE (2021).

23. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. (1982) 143:29–36. doi: 10.1148/radiology.143.1.7063747

24. Trigeorgis G, Ringeval F, Brueckner R, Marchi E, Nicolaou MA, Schuller BW, et al. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai: IEEE (2016). p. 5200–4.

25. Coppock H, Jones L, Kiskin I, Schuller BW. COVID-19 detection from audio: seven grains of salt. *Lancet Digit Health*. (2021) 3:e537–8. doi: 10.1016/S2589-7500(21)00141-2

26. Singanayagam A, Patel M, Charlett A, Bernal JL, Saliba V, Ellis J, et al. Duration of infectiousness and correlation with RT-PCR cycle threshold values in cases of COVID-19, England, January to May 2020. *Eurosurveillance*. (2020) 25:2001483. doi: 10.2807/1560-7917.ES.2020.25.32.2001483

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.