On the pragmatism of using binary classifiers over data intensive neural network classifiers for detection of COVID-19 from voice

Ankit Shah[1], *Hira Dhamyal* [1], *Yang Gao*[2], *Rita Singh*[1], *Bhiksha Raj* [1]

[1]Language Technologies Institute, Carnegie Mellon University
[2]Electrical and Computer Engineering, Carnegie Mellon University
{aps1, hyd, yanggao, rsingh, bhiksha}@andrew.cmu.edu

Abstract

Lately, there has been a global effort by multiple research groups to detect COVID-19 from voice. Different researchers use different kinds of information from the voice signal to achieve this. Various types of phonated sounds and the sound of cough and breath have all been used with varying degree of success in automated voice based COVID-19 detection apps. In this paper, we show that detecting COVID-19 from voice does not require custom made non-standard features or complicated neural network classifiers rather it can be successfully done with just standard features and simple binary classifiers. In fact, we show that the latter are not only more accurate and interpretable and also more computational efficient in that they can be run locally on small devices. We demonstrate this from a human-curated dataset collected and calibrated in clinical settings. On this dataset which comprises over 1000 speakers, a simple binary classifier is able to achieve 94% detection accuracy.

Index Terms: Detecting Covid-19 from Voice, Binary classifier, CNN for voice analysis.

1. Introduction

The Covid-19 pandemic no longer needs an introduction, having sickened – and killed – millions of people worldwide in just a few months [1]. One of the primary weapons on our arsenal to combat this pandemic is isolation – identifying infected or potentially infected individuals, and isolating them, to prevent them from infecting others. Identifying such individuals requires tests, and testing capacity for Covid-19 remains inadequate, with often unacceptable wait times during which an individual may spread the infection [2]. Consequently the search remains ongoing for rapid tests that can be quickly dispensed with minimal logistical challenges. Voice-based testing has been identified as one such possibility [3].

Covid-19 primarily manifests as an illness of the respiratory tract [4], although other organs too are affected. In symptomatic patients, it affects the lower and upper respiratory tract [5]. While, by the very widespread nature of the pandemic and the shortage of testing capacity, the actual number of infected people and fraction of this population that exhibited specific symptoms can never be known precisely, various reports suggest that between 60% and 80% of symptomatic patients develop a dry cough [6], and roughly a third develop a wet cough [6]. A common symptom is respiratory distress , and fluid accumulation in the lungs is frequently reported even in early-stage patients [7].

Naturally, it is to be expected that these impairments will manifest in the patient's vocalizations – both in the form of coughing and in changes to their voice. Covid-19 affects the three stages in voice production: respiration, phonation and the underlying resonant system [4]. The reduced lung capacity affects exhalation, affecting the mechanisms for the subglottal pressure and disrupting normal phonation [4]. In a self-assesment study, COVID patients reported difficulty producing certain voiced sounds and noticed changes in their voice [8].

Consequently, a number of research groups around the world have initiated efforts on attempting to diagnose potential Covid infections from recordings of vocalizations [9, 5]. While most groups have focused on cough sounds [10, 11, 12] as they are a frequent symptom of Covid-19, several groups have also considered other vocalizations, such as breathing sounds [10, 13] extended vowels [14, 15, 16], and counts. Yet other teams have analyzed free-form speech such as those obtainable from YouTube recordings[17]. The general consensus appears to be that cough sounds [13], and possibly speech sounds do provide discriminative information about the presence of Covid-19. Apart from this, there are a few works which explore knowledge-driven models, e.g [9, 5] have found differences in the vocal fold oscillations between Covid-19 patients and non Covid-19 people.

In all of these efforts, researchers have ubiquitously used neural networks for classification. Neural networks however require larger amounts of data for robust training. In the case of Covid-19 it is difficult to get sufficient curated data for robust neural network training. Data insufficiency has cast a doubt on generalization ability of the results obtained by previous studies. In this work, our main focus was on obtaining high amounts of medically curated data. We work directly with data curated by Merlin Inc., a private firm in Chile working directly with a hospital and the Covid-19 patients.

We explore the basic premise of whether voice samples can accurately be used to detect Covid-19 from voice. The question that this paper explores is : if we use classifiers for which data sufficiency is not such a critical issue, would the detection of Covid-19 from voice be positively affected. If this were feasible, that is if simpler and less data intensive classifiers were able to achieve good results for detecting Covid-19 from voice then we could exploit this to create much faster, more robust and more accurate and more portable Covid-19 detection applications.

In this paper, we explore several simple classifiers and compare and contrast them to neural network based classifiers using standard features available publicly for voice analysis/speech applications. The latter is a choice that has been deliberately made in order to avoid the extra effort of inventing new features in a time-critical pandemic situation. It brings out the value of confidently relying on time-tested features and classifiers for detecting novel entities from voice. It brings out the importance of doing this exercise first before presuming that the situation is complex and that it requires complex solutions. Sometimes complex challenges do have easy to implement and simple solutions. It is purpose of this paper to highlight this.

The rest of the paper is organized as follows. In Section 2, we discuss the work that has been done for detecting Covid-19 from voice from multiple perspectives. In Section 3, we describe the features and classifiers that have been used in multiple settings and explored by us in this paper for bringing out our main points. Section 4 describes the our data and experiments. In the section 5, we present our conclusions.

2. Related Work

Analysis of vocal sounds has been investigated for the detection of numerous diseases [18]. Recently many researchers have developed voice based analyzers for the detection of Covid-19 [5, 12, 16, 11, 19]. So far, the research has focused on two aspects – the collection and curation of reliable data in possibly non-clinical settings (e.g. crowsourced settings) and the determination of which vocal sounds might be most discriminative for the detection of Covid-19. Examples include Chloe et.al [10] who collect cough and breathing samples from people who self-report as being infected with Covid. They have a larger dataset but with weaker correlation between ground truth and reported status of the voice donor, compared to fully clinically validated data. They run SVM and Logistic Regression classifiers with handcrafted spectral features, but the results are unreliable due to the errors that can be inadvertently or deliberately introduced in the process of self-reporting, as mentioned above.

Other efforts include analyzing selected vowel sounds from speech recordings, e.g counting from 50 to 80, using neural networks, as in the study in [15], which analyzes data collected over cellular phones . It concludes that counting is useful in distinguishing Covid sounds, in particular the /z/ sound. [4] report results on data collected from clinical settings and descuss how various acoustic parameters of vowel sounds like /aa/ change between the infected and non-infected people. The parameters analyzed include fundamental frequency, jitter and shimmer, HNR etc.

The AI4Covid Application (described in [11]) performs the analysis on 70 samples of positive COVID-19 and 96 negative samples and reports 89% F-1 score for the detection of Covid-19. However, since the data size used is small, the authors do not perform cross-validation experiments.

The study in [15] performs analysis on 29 positive and 59 negative COVID-19 patient's data with the sounds /ah/, /z/ and also counting, all of which are recorded at the high sampling rate if 32 kHz for all speakers. Their model extracts a feature representation using a transformer RNN network architecture with an added speaker dissimilarity loss. The final score for their model is obtained using a libSVM based classifier, yielding an AUC of 0.78, obtained with a leave-one-out based cross-validation strategy, owing to the small dataset size. [12] performs analysis on a relatively large-scale dataset comprising 376 positive cough samples, and 663 negative cough samples. The analysis uses a ResNet-based CNN architecture for making a binary prediction (Covid vs non-COVID) and yields an AUC of 0.68 on their dataset. However, their method uses a relatively high portion (90%) of the total data for training.

The Coswara dataset [20] comprises breathing, cough, vocalized sounds, which include the vowels /e/, /i/, /a/, shallow and fast deep breathing sounds collected from COVID positive and negative patients. They perform classification using a 9-class classification for the sound categories, with features such as spectral contrast, MFCC, spectral roll-off, zero-crossing rate, spectral bandwidth, spectral flatness etc., and achieve an accuracy score of 66.74% on their test set. The low accuracy presumably stems from the split of the (small amount of) data and from using an unnecessarily high sampling rate (48 kHz) which causes a loss of discriminability in the classification space.

[13] has collected data from 7000 unique users with 235 positive samples only. Features extracted include standard ones such as duration, tempo, period, RMS, energy, spectral centroid and Vggish based features to generate a 477-dimensional vector from custom features and a 256-dimensional Vggish feature representation. The data collected comprises cough and breath sounds, and the approach yields an AUC of 0.8 for the COVID vs non-COVID classification task.

The research community has yet to work on a dataset with a high number of positive and negative speakers wherein the dataset consists of diverse categories of audio samples which would be an appropriate dataset size to perform k-fold cross-validation with a large-parameter neural network classifier. Our goal in this paper is to find the turning point or cusp where the need for large amounts of data is minimized which the classification accuracy is not compromised. For this, we explore standard less-data intensive classifiers to find the best balance of dataset size, features and classifiers, to achieve our goal.

3. Features and Classifiers

3.1. Features Extraction

We extract a multitude of features from voice. We use the OpenSMILE toolkit [21], using the inbuilt *emobase2010.conf* configuration file. This configuration file extract features such as intensity and loudness, cepstrum (e.g MFCC), LPC, pitch and voice quality (e.g jitter, shimmer). The OpenSMILE feature configuration file thus extracts a 1582 dimensional feature vector per audio recording. We call this Ovec feature in Section 5 We use features extracted from the pre-trained PANN model [22] on audio datasets and from the pre-trained model on VGG dataset [23]. We extract feature representations from the PASE model [24, 25] and combine them for analysis. The PASE model feature representation is extracted from the CNN block containing 64 units prior to the linear feed forward network. PASE features were generated with the dimensionality of 256 x number of frames. The PASE features were then normalized to the training dataset to generate feature representation which is scaled with zero mean and unit variance.

We extract the spectral features of audio files using librosa [26] as our audio processing library. The spectral features include,

- `Zero Crossing Rate` - The rate at which the speech signal passes the zero value.
- `Spectral Centroids` - The weighted mean of frequencies in the speech spectrogram.
- `Spectral Roll Off` - This computes the roll off frequency, the point under which 85% of the power signal is contained.
- `Tempo` - It is a measure of the beats per minute in the signal.
- `Root Mean Square Energy` - This computes RMS energy per frame in the signal based on audio samples.
- `MFCC` - Mel-Frequency Cepstral Coefficients a.k.a.the coefficients of Mel-frequency cepstrum, are one of the most commonly considered spectral features. We used the first 20 coefficients for this experiment.
- `MFCC Delta First Order` - Temporal MFCC delta features.

- `MFCC Delta Second Order` - Acceleration MFCC delta features.

For all these spectral features except Tempo, we extracted several statistical features such as min, max, mean, rms (root mean square), median, inter quartile range, first, second and third quartile, standard deviation, skew, and kurtosis, to get better representation of the data. This collectively returned a (14,1) statistical feature per spectral feature. For MFCCs we concatenate the 14 dimension spectral feature per MFCC coefficient, obtaining a 280 sized vector. Eventually, we obtain 833 spectral features which we call as the Custom feature extractor as referred in section 5. We extracted the YAMNet based feature extractor as well as Open-L3 [27] based feature extractor for each voice recording to compare their performance with Vggish [28] based feature extractor.

3.2. Classifier Description

3.2.1. CNN model with Pase and Spectrogram features

We built CNN models as our base classifier. Based on the scale of our collected dataset, one important aspect we want to have is the generalization of our algorithm. In the CNN experiments, we use two set of features: the spectrogram and the problem agnostic speech encoder (PASE) [24, 25] features. Pase features are designed to be general, robust and transferable features that capture the meaningful information of speech and less likely to contain the superficial features which were sufficient for the training data only.

3.2.2. Machine learning based binary classifiers

We use simpler machine learning based binary classifiers such as RandomForest, Support Vector machines, Logistic regression in order to perform classification.

4. Data and Experiments

4.1. Dataset and its description

In this section, we describe the collection process and the statistics of the COVID-19 voice dataset. We used a dataset collected under clinical supervision and curated by Merlin Inc., a private firm in Chile. The subjects usually have symptoms of coughing, sneezing, breathing difficulties etc. So these related symptoms were also recorded with the voice information to account for symptomatic and asymtomatic COVID-19 diagnosis. The COVID positive or negative label was indicated by the subject's lab-certified test results. Additionally metadata also contained information about preexisting conditions such as smoking preference, asthma and detailed comments about the speakers health during the dataset generation. The data samples were recorded over a smartphone, and sampled at 8khz.

The dataset consists of 421 positive cases and 989 negative cases. Each case here represents a unique speaker. To control the phonemic variation of the recordings, we asked the subjects to record their voice speaking alphabets a-z, counting from 1-20 and producing coughs. To limit the physical contact and to simulate real application scenario, subjects recorded in a quiet room environment. The total duration of the positive files is 17.5 hours and negative files is 20.5 hours with over 37 hours of total data.

4.2. Experiments

We perform numerous experiments for the detection of COVID where the input is only the voice signal. Each speaker has six voice recordings namely cough, the elongated vowels /AH/, /UW/ and /IY/, alphabet and count. In all experiments, the speakers in the training set were separate from those in the test set to ensure that the models do not inadvertently simply capture speaker identity. Thus, we split the data in k fold cross-validation sets based on speaker identity.

We run various classifiers on our data such as RandomForest (RF), Support Vector Machines (SVMs) with Radial Basis Function and Logistic Regression (LR). We analysed all the spectral features on each of these classifiers including combination of features like Ovec and Custom. The RandomForest runs on k fold equal to 3, 5 and 10. We run grid search on all these classifiers to find the best parameters, particularly experimenting with γ. We report the performance in the Section 5.

5. Results and Discussion

From the dataset consisting of a total of 1410 speakers (or 8460 audio recordings), the entire metadata was received for a total of 815 speakers, out of which there were 296 males and 519 females.

5.1. Age based Analysis

We have broadly categorized the ages into four groups, including the ones whose age was missing, shown in Table 1.

- Group 1 : age $<=$ 30 years
- Group 2 : 30 $<$ age $<=$ 40 years
- Group 3 : 40 years $<$ age
- Group 4 : Age was missing

For patients with age $>$ 40, our classifier is capable of recognizing COVID-19 patients accurately 65% of the times whereas those with age $<$ 30, classifier is capable of accurately detecting 56% of the times. The classifier is more reliably able to detect COVID-19 in case of females over males.

Gender Group1	Group2	Group3	Group4	
Male	65	80	80	71
Female	127	129	145	115

Table 1: *Gender based age analysis*

We find Sore-throat is present in most of COVID-19 cases where we are able to detect COVID-19 in 57% of the cases. When sneeze is present, the classifier is capable of detecting COVID-19 in 47% of the cases. When cough is a known symptom for the speaker, then our classifier detects COVID-19 in 73% of the cases.

5.2. Smokers and People with Asthma

Having a previous history of asthma or smoking might put people at a higher risk of COVID [29]. We received data where people had a history of asthma and/or were smokers. Table 2 shows the statistics in our data. Our experimental results indicate asthama is more correlated to COVID-19 than smoking. For the patients with COVID-19 and smoking habit, our classifier is accurate by detecting COVID-19 in 66 % of cases whereas those with COVID-19 and asthama cases we are able to detect COVID-19 accurately in 80 % of the cases.

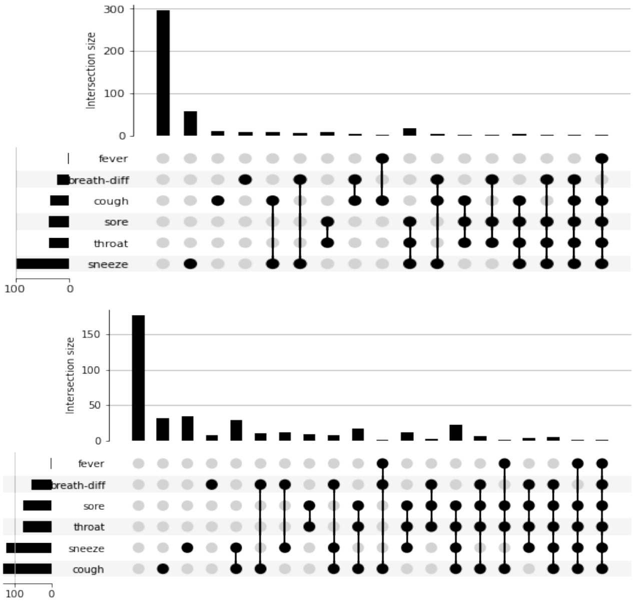

Figure 1: *The top figures shows the distribution of symptoms in COVID-19 positive patients and the lower figure is the distribution of symptoms in COVID-19 negative individuals (breath-diff stands for breathing difficulty). The rows represents each symptom whereas the column represents the frequency of occurrence of a symptom which are marked by dots*

Gender	Has Asthma	Is Smoker
Male	12	95
Female	38	139

Table 2: *Gender based asthma and smoker statistics*

5.3. Difference in days of diagnosis and data collection

The data collection process introduces a delay between the COVID onset and when the voice sample was recorded. Each speaker is labelled for the day they got diagnosed and the day they recorded their data. Table 3 shows number of people distribution depending on when they were diagnosed (eg. 1 week ago, between 1-2 weeks or greater than 2 weeks)

Diagnosis Day Range	Number of People
Diagnosis day <= 7 days	95
7 days < Diagnosis day <= 14 days	209
Diagnosis day > 14 days	159
Diagnosis day not known	37

Table 3: *Difference between Diagnosis Day and Data collection day*

In general, we find that COVID-19 can be more reliably detected for patients whose voice samples have been collected within 14 days of diagnosis, than for those whose samples were collected after this period.

5.4. Audio type analysis

We analyze classifier performance according to the audio-type, to identify which type of recording is most suited to capture the voice signatures related to COVID-19. Our results indicate that vowel /IY/ and vowel /UW/ are better at detecting COVID-19 than other types of audio samples. Cough samples are also useful in detecting COVID-19.

Classifier	Feature	AUC
Random Forest	Ovec	0.76
Random Forest	PANN	0.78
Random Forest	vggish	0.59
Random Forest	Open-L3	0.64
Random Forest	YAMNet	0.72
Random Forest	Custom Feat	0.82
Random Forest	Ovec + Custom Feat	0.84
Random Forest	Ovec + Custom Feat + vggish	0.82
Random Forest	Ovec + custom feature + YAMNet	0.86
CNN	Spectrogram	0.69
CNN	PASE	0.73

Table 4: *Results obtained on the full dataset set. Here Ovec refers to features obtained using OpenSMILE, PASE using PASE architecture, custom spectral feature. Random Forest classifier is robust and have improved performance in comparison to CNN based classifier*

5.5. Overall Classifier results

Table 4 proves our hypothesis that a binary classifier such as Random Forest can perform better for COVID-19 detection than CNN counterparts when faced with limited data. Our experiments have observed that Open-L3 features perform the same as vggish feature representations. Figure 2 shows the best AUC score achieved of 0.94 and ROC score of 0.85 for the Random forest based classifier with Ovec and custom features on a trimmed down dataset where 20% samples were heard to contain noise in the recordings.

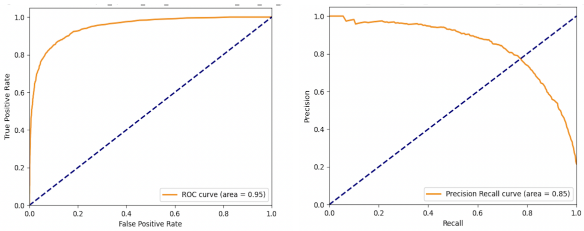

Figure 2: *AUC of 0.94 and ROC of 0.85 is obtained using Random Forest based classifier with Ovec and custom feature vector*

6. Conclusion

Motivated by the urgent need to have alternative methods to augment the medical tests, we designed the experimental setup for the COVID-19 analysis using a wide variety of voice samples. Our analysis provided insights into the type of voice sample which would work best for detecting COVID-19 patients which is the eee and ooo sounds. Our COVID-19 dataset used in the study was acquired through self-recording using a smartphone application making acquistion feasible at large scale. The preliminary results imply a feasibility for the use of this globally accessible data collection for Sars-COV-2 detection although it doesn't replace RT-PCR or RAT tests. Our results prove that using a binary classifiers such as Random Forest are more feasible to separate between COVID-19 vs non COVID-19 speakers over data intensive neural network classifiers. We find these binary classifiers to generate better and robust results given the limited amount of data available for analysis.

7. References

[1] Worldometer, "https://www.worldometers.info/coronavirus/," 2020. [Online]. Available: https://www.worldometers.info/coronavirus/

[2] H. Resources and S. Administration, "Health center covid-19 survey testing," 2020. [Online]. Available: https://bphc.hrsa.gov/emergency-response/coronavirus-health-center-data

[3] B. Insider, "Do i sound sick to you? researchers are building ai that would diagnose covid-19 by listening to people talk." 2020. [Online]. Available: https://www.businessinsider.com/ai-labs-diagnose-covid-19-voice-listening-talk-2020-4

[4] M. Asiaee, A. Vahedian-azimi, S. S. Atashi, A. Keramatfar, and M. Nourbakhsh, "Voice quality evaluation in patients with covid-19: An acoustic analysis," *Journal of Voice*, 2020.

[5] M. Al Ismail, S. Deshmukh, and R. Singh, "Detection of covid-19 through the analysis of vocal fold oscillations," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 1035–1039.

[6] W.-J. Song, C. K. Hui, J. H. Hull, S. S. Birring, L. McGarvey, S. B. Mazzone, and K. F. Chung, "Confronting covid-19-associated cough and the post-covid syndrome: role of viral neurotropism, neuroinflammation, and neuroimmune responses," *The Lancet Respiratory Medicine*, vol. 9, no. 5, pp. 533–544, 2021.

[7] C. Suess and R. Hausmann, "Gross and histopathological pulmonary findings in a covid-19 associated death during self-isolation," *International journal of legal medicine*, vol. 134, no. 4, pp. 1285–1290, 2020.

[8] J. R. Lechien, C. M. Chiesa-Estomba, P. Cabaraux, Q. Mat, K. Huet, B. Harmegnies, M. Horoi, S. D. Le Bon, A. Rodriguez, D. Dequanter *et al.*, "Features of mild-to-moderate covid-19 patients with dysphonia." *Journal of Voice*, 2020.

[9] S. Deshmukh, M. Al Ismail, and R. Singh, "Interpreting glottal flow dynamics for detecting covid-19 from voice," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 1055–1059.

[10] C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo, "Exploring automatic diagnosis of covid-19 from crowdsourced respiratory sound data," *arXiv preprint arXiv:2006.05919*, 2020.

[11] A. Imran, I. Posokhova, H. N. Qureshi, U. Masood, M. S. Riaz, K. Ali, C. N. John, M. I. Hussain, and M. Nabeel, "Ai4covid-19: Ai enabled preliminary diagnosis for covid-19 from cough samples via an app," *Informatics in Medicine Unlocked*, vol. 20, p. 100378, 2020.

[12] P. Bagad, A. Dalmia, J. Doshi, A. Nagrani, P. Bhamare, A. Mahale, S. Rane, N. Agarwal, and R. Panicker, "Cough against covid: Evidence of covid-19 signature in cough sounds," *arXiv preprint arXiv:2009.08790*, 2020.

[13] C. Brown, J. Chauhan, A. Grammenos, J. Han, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo, "Exploring automatic diagnosis of covid-19 from crowdsourced respiratory sound data," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 3474–3484. [Online]. Available: https://doi.org/10.1145/3394486.3412865

[14] R. Singh, *Voice Profiling Technologies and the detection of Covid-19 from Voice*, 2020 (accessed October 23, 2020). [Online]. Available: https://health-sounds.cl.cam.ac.uk/workshop20/rita_singh.mp4

[15] G. Pinkas, Y. Karny, A. Malachi, G. Barkai, G. Bachar, and V. Aharonson, "Sars-cov-2 detection from voice," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 1, pp. 268–274, 2020.

[16] B. W. Schuller, D. M. Schuller, K. Qian, J. Liu, H. Zheng, and X. Li, "Covid-19 and computer audition: An overview on what speech & soundanalysis could contribute in thesars-cov-2 corona crisis," *Frontiers in digital health*, vol. 3, p. 14, 2021.

[17] A. Shukla, "Covid-19 pandemic: An analysis of popular youtube videos as an alternative health information platform," *Health Informatics Journal*, vol. 27, no. 2, p. 1460458221994878, 2021.

[18] R. Singh, A. Shah, and H. Dhamyal, "An overview of techniques for biomarker discovery in voice signal," *arXiv preprint arXiv:2110.04678*, 2021.

[19] V. Despotovic, M. Ismael, M. Cornil, R. Mc Call, and G. Fagherazzi, "Detection of covid-19 from voice, cough and breathing patterns: Dataset and preliminary results," *Computers in Biology and Medicine*, vol. 138, p. 104944, 2021.

[20] N. Sharma, P. Krishnan, R. Kumar, S. Ramoji, S. R. Chetupalli, P. K. Ghosh, S. Ganapathy *et al.*, "Coswara–a database of breathing, cough, and voice sounds for covid-19 diagnosis," *arXiv preprint arXiv:2005.10548*, 2020.

[21] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[22] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *arXiv preprint arXiv:1912.10211*, 2019.

[23] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017. [Online]. Available: https://arxiv.org/abs/1609.09430

[24] S. Pascual, M. Ravanelli, J. Serrà, A. Bonafonte, and Y. Bengio, "Learning problem-agnostic speech representations from multiple self-supervised tasks," *arXiv preprint arXiv:1904.03416*, 2019.

[25] M. Ravanelli, J. Zhong, S. Pascual, P. Swietojanski, J. Monteiro, J. Trmal, and Y. Bengio, "Multi-task self-supervised learning for robust speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6989–6993.

[26] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015.

[27] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen and learn more: Design choices for deep audio embeddings," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, May 2019, pp. 3852–3856.

[28] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.

[29] N. C. for Immunization and D. o. V. D. Respiratory Diseases (NCIRD), "People with moderate to severe asthma," 2020. [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/asthma.html