ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746015

DETECTION OF COVID-19 FROM JOINT TIME AND FREQUENCY ANALYSIS OF SPEECH, BREATHING AND COUGH AUDIO

John Harvill[1], *Yash Wani*[2], *Moitreya Chatterjee*[1], *Mustafa Alam*[2], *David G. Beiser*[2], *David Chestek*[3], *Mark Hasegawa-Johnson*[1], *Narendra Ahuja*[1]

[1]University of Illinois at Urbana-Champaign, United States
[2]University of Chicago, United States
[3]University of Illinois at Chicago, United States

ABSTRACT

The distinct cough sounds produced by a variety of respiratory diseases suggest the potential for the development of a new class of audio bio-markers for the detection of COVID-19. Accurate audio biomarker-based COVID-19 tests would be inexpensive, readily scalable, and non-invasive. Audio biomarker screening could also be utilized in resource-limited settings prior to traditional diagnostic testing. Here we explore the possibility of leveraging three audio modalities: cough, breathing, and speech to determine COVID-19 status. We train a separate neural classification system on each modality, as well as a fused classification system on all three modalities together. Ablation studies are performed to understand the relationship between individual and collective performance of the modalities. Additionally, we analyze the extent to which temporal and spectral features contribute to COVID-19 status information contained in the audio signals.

Index Terms— DiCOVA-II, COVID-19, Telemedicine

1. INTRODUCTION

The COVID-19 pandemic has impacted many countries globally. High-frequency screening with rapid antigen tests has been recommended by the Center for Disease Control as one strategy for mitigating the spread of COVID-19 in high-density communities and critical infrastructure workplaces [1]. While such strategies have been shown to be feasible in small, resource rich settings, such as college campuses, private hospitals, and industrial workplaces, they can be expensive and logistically difficult to implement at large scale. Furthermore, implementation obstacles can become insurmountable in lower-middle income countries (LMIC), disadvantaged communities with historically poor access to care, and rural populations who rely on remote telemedicine care [2]. Thus, the development of rapid, low-cost, widely scalable COVID-19 screening tools could significantly contribute to the global effort to control the COVID-19 pandemic [3].

The SARS-CoV-2 virus gains access to the body through specific binding within the respiratory epithelium. Accordingly, some of the earliest and most pronounced symptoms of COVID-19 involve the airway with common respiratory symptoms including nasal congestion, sore throat, cough, and shortness of breath arising within 2-14 days of exposure [4]. While such symptoms are common to a variety of upper and lower respiratory infections due to, for example, influenza virus, para-influenza virus, respiratory syncytial virus (RSV), rhinovirus, adenovirus, and bordatella pertussis, there is evidence to suggest that certain respiratory disease entities have clinically-relevant acoustic signatures. For example, infectious croup, a disease most commonly caused by parainfluenza virus in children causes swelling of mid to large airways often resulting in a distinctive barking cough and a harsh inspiratory sound known as stridor [5].

By contrast, RSV produces inflammation of the lower airways resulting in a characteristic pattern of crackles, wheezing, and prolonged expiration in infants [6]. Perhaps the most historically well-known clinical presentation is that of Whooping cough caused by Bordetella pertussis, which produces uncontrollable, violent coughing bouts [7]. Together, these examples support the feasibility of audio-based respiratory disease classifiers. We further hypothesize that it will be possible to develop an accurate audio-based COVID-19 disease classifier that utilizes data from a patient's speech, breathing, and cough patterns.

2. RELATED WORK

Significant past research in audio-based respiratory disease classification has enabled rapid progress in remote COVID-19 prediction from audio signals alone [8, 9, 10, 11]. For example, Laguarta et al. [12] used a poisson biomarker layer and three ResNet-50 CNNs in parallel to output a binary diagnostic label given MFCC features calculated from cell phone recordings of coughs. Pahar et al. [13] tried a set of seven machine learning classifiers and found that a CNN architecture was best able to discriminate between COVID-19 positive and healthy coughs while an LSTM was best able to distinguish between COVID-19 positive and negative coughs. The 2021 Spring Diagnosing COVID-19 Using Acoustics (DiCOVA) challenge aimed to build off much of this past work and featured 29 participant teams, including ours, using cough signals alone to predict the COVID-19 status of an individual. We placed third in the original DiCOVA challenge using our choice of model architecture, pre-training method, and data augmentation [14].

Several groups have attempted to leverage information from multiple audio modalities (i.e. cough, breathing, speech) to detect COVID-19. Lella et al. [11] incorporated information from cough, breathing, and speech data by using a multi-channel deep convolutional neural network trained on denoised spectrograms, Gammatone Frequency Cepstral Coefficients filter banks and Improved Mel-frequency Cepstral Coefficients. Cough, breathing, and speech log Mel spectrograms were concatenated prior to being fed into the model. Coppock et al. [12] used a ResNet architecture to detect COVID-19 using breathing and cough audio. Similar to [11], spectrogram features were combined and provided to the model

This work was supported by the C3.ai Digital Transformation Institute.

Authorized licensed use limited to: East China Normal University. Downloaded on September 20,2022 at 15:37:29 UTC from IEEE Xplore. Restrictions apply.

together.

A variety of features have been used in audio-based classification of diseases like COVID-19 [15]. Time-frequency features including spectrograms, log Mel spectrograms, and MFCCs have become popularized by their ability to capture information from the audio signal in both the time and frequency domains [16]. Features that capture only temporal variation have also been found to be important as predictive features. For example, Breebaart et al. [17] showed that auditory filterbank temporal envelopes were able to outperform MFCC features in the differentiation of five audio classes. Presumably, pathophysiological information is differently available in temporal and spectral features, and may perhaps even be distributed differently between each modality: cough, breathing, and speech. We perform ablation studies to explore this relationship.

3. DATA

Our approach relies on pretraining for speech, cough, and breathing modalities individually, so we seek datasets with a large number of samples without the requirement of relevant labels for COVID-19 classification. For breathing, we pretrain on a subset of the COVID-19 Sounds dataset [18]. The dataset is not freely available to the public, but a subset can be obtained with a license agreement from the collectors of the dataset. The subset contains a total of 897 breathing audio clips sampled at 48kHz. We pretrained our cough system on the COUGHVID dataset [19], a collection of crowd-sourced cough sounds collected from individuals around the world via a web interface. Web recordings were passed to a cough identification algorithm and valid samples were annotated by experts. Of the over 20,000 cough recordings, 1,155 claimed to be COVID-19 positive. All audio were sampled at 48 kHz and no associated metadata was used. Our speech model was pretrained on the train-clean-100 subset of the Librispeech dataset [20], constituting approximately 100 hours of English read speech sampled at 16kHz.

Fine-tuning was performed on the DiCOVA-II dataset [21], a collection of 965 audio files each for breathing, cough, and speech. The dataset had 172 COVID-19 positive samples and 793 COVID-19 negative samples. Patient gender and COVID-19 status were the only metadata provided and audio files were sampled at 44.1 kHz. Included with the training data is a five-fold cross-validation split with approximately 200 samples held out for validation in each fold. The blind test data used for the leaderboard system rankings consists of 471 audio files per modality. Gender metadata was provided, but the ground-truth COVID-19 labels were not included.

Preprocessing: Given that the DiCOVA-II dataset is sampled at 44.1kHz and the possibility of important information being contained in higher frequencies, we first resample all audio recordings to 40kHz (opposed to a standard 16kHz for speech processing). We then compute 160-dimensional log Mel spectrograms with a window of 2048 samples and step size of 400 (10ms). The spectrograms are thresholded, setting any components less than -120dB to -120dB. The spectrograms are then normalized to between 0 (previously -120dB) and 1 (previously 0dB).

4. METHODS

Baselines: The DiCOVA-II challenge organizers provide all competitors with a single baseline. Log Mel spectrograms (64 dimensional + delta + delta-delta) with a window size of 1102 and hop length of 441 are used as feature vectors. The authors use a bidirectional long short term memory network (BLSTM) to classify the samples as COVID-19 positive or negative. The network has two encoder layers with 128 neurons each, dropout of 0.1, and average pooling followed by a classifier with 64 neurons with a dropout of 0.1 and a tanh activation function. The baseline fusion approach simply takes the mean of the probabilities from each modality.

Pretraining: Due to the relatively small amount of training data, we hope to avoid overfitting by learning modality-specific features on a large amount of unlabeled data during the pretraining stage. We use Autoregressive Predictive Coding (APC) [22], a pretraining technique for audio that seeks to predict future input spectral frames given past frames. We choose to use 10 future frames and the output from lower layers in the pretraining network, because these were the optimal hyperparameters we found in our previous work for determination of presence of COVID-19 from cough sounds [14]. We use the same model configuration from our previous work as well. We choose the model checkpoint based on the best validation loss after the validation loss flattens on a log scale.

Finetuning: We finetune in the same manner as in our previous work [14]. We take the output from the frozen lower layers of our pretrained model per modality and stack an additional two Bidirectional Long Short Term Memory (BLSTM) layers. We take the final states of the forward and backward directions, concatenate and then pass through three more fully-connected layers. We take the softmax at the output to predict the probability of the sample being COVID-positive. We choose model checkpoints based on the AUC on the validation data, train with a batch size of three, and use a learning rate of 0.0001.

Fusion: To combine the knowledge obtained individually from each modality, we want to train a small neural network to combine the predictions from cough, speech and breathing in an optimal way. Given the three modality pretrained and finetuned networks, we pass in audio samples and take the last hidden state of the fully-connected layer from each finetuned model and concatenate to get a vector representing cough, speech and breathing information for the patient. We then pass this vector through a small three-layer fully-connected network for the final prediction (see Figure 1). We train on all five folds and take the mean of the probabilities produced by the systems trained on each fold as an ensemble for the test predictions. Unlike during finetuning, we use a much larger batch size of 50 and choose the final models based on validation loss instead of AUC. We find that while the ensembled validation AUC is better when choosing models based on validation AUC, the test AUC is lower.

Augmentation and Metadata: As we found in our previous work [14], avoiding overfitting with such a small dataset is critical for improved performance. We use two forms of data augmentation, the first of which is SpecAugment [23]. We find empirically that masking improves performance but time warping does not. The second form of augmentation is random cropping. The motivation for this augmentation is threefold (1) increase the relative number of training examples (2) reduce training time and (3) similar approaches have been successful in computer vision [24]. Random cropping consists of randomly selecting a contiguous section of length $x \cdot L$ from the original spectrogram as input where L is the original length of the spectrogram. We choose $x = 0.85$, $x = 0.3$, and $x = 0.3$ for cough, breathing and speech, respectively. While gender (male/female) metadata are provided for training and test data, we find that using this information during finetuning in the form of an additional embedding concatenated to the end of the framewise hidden states output by the pretrained model reduces overall performance. We believe this is due to overfitting, and thus do not use gender labels in our final system.

Authorized licensed use limited to: East China Normal University. Downloaded on September 20,2022 at 15:37:29 UTC from IEEE Xplore. Restrictions apply.

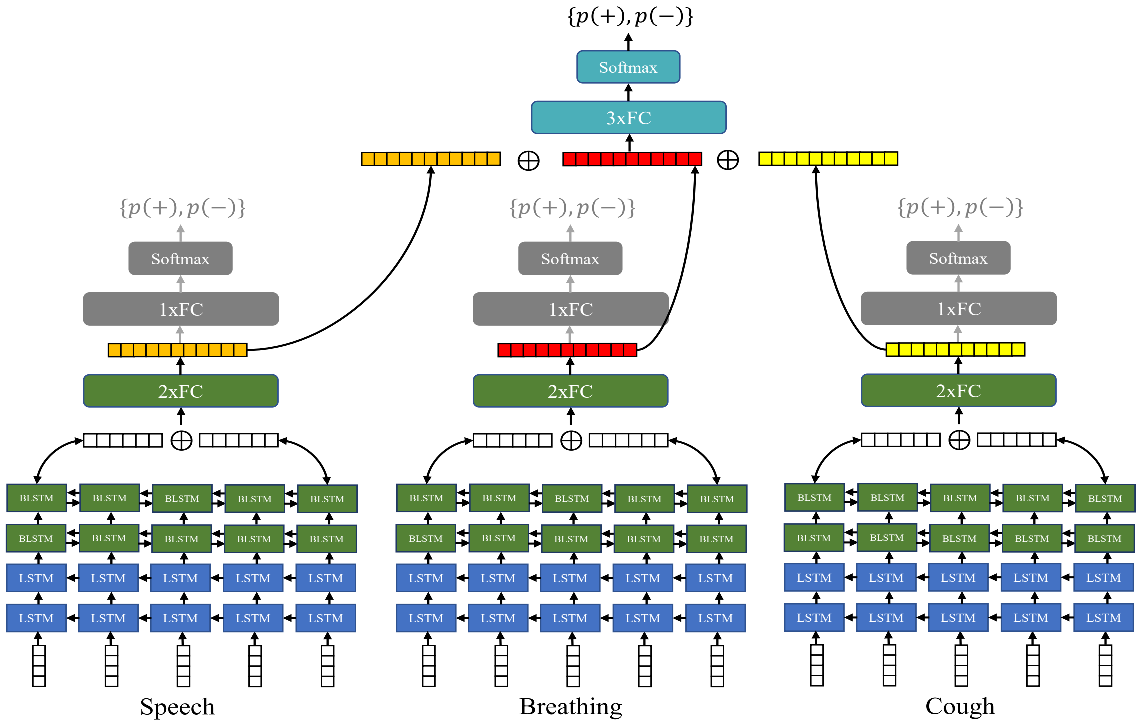

Fig. 1. Fusion System Overview: Individual modality models are first trained using pretraining and finetuning, then hidden representations are concatenated for training of the fusion network. Blue layers are frozen after pretraining, green layers are frozen after finetuning, gray components are unused after finetuning, and teal layers are the only parameters updated during fusion training.

5. EXPERIMENTS

	S	B	C	F
Test				
Baseline	84.26	84.50	74.89	84.70
Proposed	84.73	85.77	70.34	**87.26**
Validation				
	S	B	C	F
Baseline	80.16	77.25	75.21	81.67
Proposed	81.51	78.68	72.20	**85.79**

Table 1. Comparison of baseline and proposed approach for each modality type on both test and validation data using AUC as evaluation metric. Scores are copied from the DiCOVA-II challenge leaderboard. Modalities are indicated as: Speech=S, Breathing=B, Cough=C, Fusion=F. The best score for test/validation data is bolded. Baseline scores are from team "dicova" on the leaderboard.

	$T = 0.25$		$T = 0.5$		$T = 0.75$	
	F1	Agr %	F1	Agr %	F1	Agr %
S+B+C	**0.72**	28	**0.76**	52	0.21	74
S+B	0.63	61	0.66	72	**0.51**	81
S+C	0.64	41	0.65	66	0.2	85
C+B	0.47	56	0.6	67	0.3	83
S	0.52	100	0.51	100	0.45	100
B	0.41	100	0.49	100	0.49	100
C	0.35	100	0.43	100	0.23	100
Mean	0.53	69	0.59	80	0.34	89

Table 2. Comparison of F1 scores for different agreement sets across modalities on validation data. Similar to Table 1, Speech=S, Breathing=B, Cough=C. Presence of multiple modalities indicates that the modalities agreed on the predicted label at the evaluation threshold. "Agr %" represents the percentage of samples for which the modalities agreed on the predicted label. Largest F1 score per threshold is bolded.

The DiCOVA-II challenge has four tracks which are (1) Breathing (2) Cough (3) Speech (4) Fusion. Each track indicates the modality that can be used for prediction, where the Fusion track allows use of all three modalities. Our code is available at `https://github.com/jharvill23/Fa21DiCOVA-II`.

We train our proposed pretraining + finetuning system for all five folds on each individual modality track. We then train our proposed fusion system for all five folds on the fusion track. Average validation performance on the five folds and test performance on the 471 held-out blind test samples are reported on the DiCOVA-II challenge leaderboard based on reported sample probabilities output from our model. We compare performance between that of our proposed approach and the DiCOVA-II baseline method in Table 1. In order to explore the importance of each modality in more depth, we show additional results using our proposed approach where agreement between different modalities is considered in Table 2. For each modality combination, we report F1 at classification thresholds of $T = 0.25$, $T = 0.50$, and $T = 0.75$ on the subset of samples where the modalities considered agree on the classification label. For example, when looking at the speech and breathing combination (S+B), samples that the speech model classifies as positive and breathing model classifies as negative (and vice versa) will be ex-

Authorized licensed use limited to: East China Normal University. Downloaded on September 20,2022 at 15:37:29 UTC from IEEE Xplore. Restrictions apply.

cluded. Our motivation to show these results is to indicate that when different modalities are in agreement, confidence in the prediction can be increased. In the cases where it occurs, this can improve usability in a practical implementation setting.

6. RESULTS

In Table 1 we can see that our proposed approach improves over the baseline for speech and breathing modalities on the held out test data. Most importantly, our fusion system outperforms the baseline by several points. We also note that we place second in the DiCOVA-II challenge on the leaderboard for the fusion track (user jharvill23 at `https://competitions.codalab.org/competitions/34801#results`), where the best performance improves over our approach by less than 1.5 AUC points on the test data. We apply the same models to each modality, but find that performance is worse than the baseline for cough. We noticed similar trends on the DiCOVA-II challenge leaderboard, where teams ranked higher or lower for different modalities. This suggests that future work could be devoted to tailoring approaches to modalities individually, since one blanket approach does not appear to work best.

In Table 2 we can see results of our system evaluated on subsets of data where different modalities agree on the label given a certain threshold. First, notice that the best performance at each threshold is generally achieved when more modalities agree on the label. There is a downward trend in performance as less modalities are required to agree for thresholds of $T = 0.25$ and $T = 0.5$ with varying performance for $T = 0.75$. Second, notice that the percentage of samples where modalities agree decreases as more modalities are added. For a well-chosen threshold ($T = 0.5$), agreement between more modalities is less likely, but when it occurs, confidence in the prediction is increased.

7. ABLATIONS

Spectrogram vs. Cumulative Amplitude Input: We hypothesize that each modality may contain different amounts of information with respect to timing or spectral content. We empirically address this question by comparing performance of our proposed approach (without pretraining) using either the spectrogram or a collapsed version of the spectrogram that we denote as "cumulative amplitude". Given a sequence of log Mel filterbank feature vectors $x_1, x_2, ..., x_T$ where T is the length of the input sequence and the dimension of each x_t is N (number of Mel frequency bins), the "cumulative amplitude" feature e_t at each timestep is:

$$e_t = \frac{1}{N}\sum_{i=1}^{N} x_{t,i} \tag{1}$$

Thus we remove all frequency relationships by collapsing each spectral frame to a scalar value indicating only amplitude of the signal at each timestep. We run the experiments for all three modalities and report the average best AUC on the validation data for all five folds in Table 3.

Given that an AUC of 50 represents random guessing, it is clear that timing information in speech is not very useful for determining COVID-19 status. This indicates that most information in speech comes from the spectral content alone. On the other hand, the timing information in both breathing and cough signals appears useful, indicating that this information may play a large role in classification even when spectrogram input is provided as in our proposed system.

	Spectrogram	Amplitude	Difference
Speech	81.99	58.85	+23.14
Cough	76.53	69.19	+7.34
Breathing	80.67	73.68	+6.99

Table 3. Comparison of average best AUC scores across all five folds when using spectrogram or cumulative amplitude ("Amplitude" above) as input to the model. "Difference" shows the difference in performance (Spectrogram - Amplitude).

Naive Bayes	Log. Regression	SVM	Random Forest
72.0	**73.4**	70.7	70.1

Table 4. Comparison of average best AUC scores across five folds when using clinical features for predicting COVID status of patients. We compare naive Bayes [25], logistic regression [26], Support Vector Machine (SVM) with a radial basis function kernel [26] and random forest algorithms applied to the data.

Clincical Features for Cough: We also examine the ability of acoustic cues in cough signals to be used as features for determination of COVID-19 status. We first choose 20 samples randomly from the DiCOVA-II dataset and have them annotated for five cues by clinicians. The cues are (i) Dryness (ii) Wetness/Productivity (iii) Wheezing (iv) Short (duration) (v) Long (duration). We then pretrain a classifier using these annotations and produce probability scores for each of the five cues for all training/validation samples in the DiCOVA-II dataset. We train four classifiers using these five probability scores per sample as input features. Results of five-fold cross validation using AUC as the evalutation metric are provided in Table 4. We empirically find logistic regression to be most effective.

8. CONCLUSIONS

In this paper we have proposed a neural system that relies on speech, cough, and breathing samples for improved determination of COVID-19 status over systems that rely on one modality alone. We compare performance of our system to the baseline provided by the DiCOVA-II challenge organizers and demonstrate significant performance improvement for the fusion track. We perform additional analyses to find that when models from each modality agree on the prediction label, confidence in the prediction is increased. This knowledge can help healthcare workers make better use of COVID-19 diagnosis systems that rely on audio, reducing healthcare costs and inefficiencies. We also demonstrate empirically that cough and breathing signals appear to contain a lot of information relevant to COVID-19 diagnosis from timing alone, because performance of the classifiers relying on cumulative amplitude input is similar to those relying on log Mel spectrogram input for these modalities. We find that information on the timing of full-band energy from speech is not of importance for determintation of COVID-19, and that the rich information contained in the speech signal comes mostly from the spectral content. We also find that systems using acoustic cues annotated by clinicians can provide a simple but rich feature set from cough sounds for COVID-19 diagnosis. In addition to the high-performing neural architecture we have proposed, the additional insights we provide into the information contained in speech, cough, and breathing signals for the purpose of COVID-19 diagnosis can motivate new design choices to solve the problem of determining COVID-19 status from audio.

Authorized licensed use limited to: East China Normal University. Downloaded on September 20,2022 at 15:37:29 UTC from IEEE Xplore. Restrictions apply.

9. REFERENCES

[1] Centers for Disease Control and Prevention, "Testing Strategy for Coronavirus (COVID-19) in High-Density Critical Infrastructure Workplaces after a COVID-19 Case Is Identified," 2020.

[2] Devon E. McMahonid, Gregory A. Peters, Louise C. Iversid, and Esther E. Freemanid, "Global resource shortages during covid-19: Bad news for low-income countries," *PLoS Neglected Tropical Diseases*, vol. 14, no. 7, pp. 1–3, 2020.

[3] 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.

[4] Joseph R. Larsen, Margaret R. Martin, John D. Martin, Peter Kuhn, and James B. Hicks, "Modeling the Onset of Symptoms of COVID-19," *Frontiers in Public Health*, vol. 8, no. August, 2020.

[5] CL Bjornson and DW Johnson, "Croup," *Lancet*, vol. 371, pp. 329–39, 2008.

[6] Andrea T. Borchers, Christopher Chang, M. Eric Gershwin, and Laurel J. Gershwin, "Respiratory syncytial virus - A comprehensive review," *Clinical Reviews in Allergy and Immunology*, vol. 45, no. 3, pp. 331–379, 2013.

[7] Tara B Spector and Eileen K Maziarz, "Pertussis," *Medical Clinics of North America*, vol. 97, no. 4, pp. 537–552, 2013.

[8] Yoonjoo Kim, YunKyong Hyon, Sung Soo Jung, Sunju Lee, Geon Yoo, Chaeuk Chung, and Taeyoung Ha, "Respiratory sound classification for crackles, wheezes, and rhonchi in the clinical field using deep learning," *Scientific Reports*, vol. 11, no. 1, pp. 1–11, 2021.

[9] Fatih Demir, Abdulkadir Sengur, and Varun Bajaj, "Convolutional neural networks based efficient approach for classification of lung diseases," *Health information science and systems*, vol. 8, no. 1, pp. 1–8, 2020.

[10] Renard Xaviero Adhi Pramono, Syed Anas Imtiaz, and Esther Rodriguez-Villegas, "A cough-based algorithm for automatic diagnosis of pertussis," *PloS one*, vol. 11, no. 9, pp. e0162128, 2016.

[11] Roneel V Sharan, Shlomo Berkovsky, David Fraile Navarro, Hao Xiong, and Adam Jaffe, "Detecting pertussis in the pediatric population using respiratory sound events and cnn," *Biomedical Signal Processing and Control*, vol. 68, pp. 102722, 2021.

[12] 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.

[13] Madhurananda Pahar, Marisa Klopper, Robin Warren, and Thomas Niesler, "Covid-19 cough classification using machine learning and global smartphone recordings," *Computers in Biology and Medicine*, p. 104572, 2021.

[14] John Harvill, Yash R Wani, Mark Hasegawa-Johnson, Narendra Ahuja, David Beiser, and David Chestek, "Classification of covid-19 from cough using autoregressive predictive coding pretraining and spectral data augmentation," *Proc. Interspeech 2021*, pp. 926–930, 2021.

[15] Francesc Alías, Joan Claudi Socoró, and Xavier Sevillano, "A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds," *Applied Sciences*, vol. 6, no. 5, pp. 143, 2016.

[16] Karthikeyan Umapathy, Behnaz Ghoraani, and Sridhar Krishnan, "Audio signal processing using time-frequency approaches: coding, classification, fingerprinting, and watermarking," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, pp. 1–28, 2010.

[17] Jeroen Breebaart and Martin F McKinney, "Features for audio classification," in *Algorithms in Ambient Intelligence*, pp. 113–129. Springer, 2004.

[18] 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," *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, Jul 2020.

[19] 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, Jun 2021.

[20] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[21] Neeraj Kumar Sharma, Srikanth Raj Chetupalli, Debarpan Bhattacharya, Debottam Dutta, Pravin Mote, and Sriram Ganapathy, "The second dicova challenge: Dataset and performance analysis for covid-19 diagnosis using acoustics," 2021.

[22] Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass, "An unsupervised autoregressive model for speech representation learning," *arXiv preprint arXiv:1904.03240*, 2019.

[23] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *Interspeech 2019*, Sep 2019.

[24] Ryo Takahashi, Takashi Matsubara, and Kuniaki Uehara, "Data augmentation using random image cropping and patching for deep cnns," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 9, pp. 2917–2931, Sep 2020.

[25] Andrew McCallum, Kamal Nigam, et al., "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*. Citeseer, 1998, vol. 752, pp. 41–48.

[26] Christopher M Bishop, "Pattern recognition and machine learning," .

Authorized licensed use limited to: East China Normal University. Downloaded on September 20,2022 at 15:37:29 UTC from IEEE Xplore. Restrictions apply.