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/ICASSP43922.2022.9746271

MODELING OF PRE-TRAINED NEURAL NETWORK EMBEDDINGS LEARNED FROM RAW WAVEFORM FOR COVID-19 INFECTION DETECTION

Zohreh Mostaani[1,2] *RaviShankar Prasad* [1] *Bogdan Vlasenko*[1] *Mathew Magimai-Doss*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2] Ecole polytechnique fédérale de Lausanne, Lausanne, Switzerland

ABSTRACT

COVID-19 is a respiratory system disorder that can disrupt the function of lungs. Effects of dysfunctional respiratory mechanism can reflect upon other modalities which function in close coupling. Audio signals result from modulation of respiration through speech production system, and hence acoustic information can be modeled for detection of COVID-19. In that direction, this paper is addressing the second DiCOVA challenge that deals with COVID-19 detection based on speech, cough and breathing. We investigate modeling of (a) ComParE LLD representations derived at frame- and turn-level resolutions and (b) neural representations obtained from pre-trained neural networks trained to recognize phones and estimate breathing patterns. On Track 1, the ComParE LLD representations yield a best performance of 78.05% area under the curve (AUC). Experimental studies on Track 2 and Track 3 demonstrate that neural representations tend to yield better detection than ComParE LLD representations. Late fusion of different utterance level representations of neural embeddings yielded a best performance of 80.64% AUC.

Index Terms— COVID-19 identification, breathing pattern estimation, phoneme recognition, ComParE features, BoAW

1. INTRODUCTION

Corona virus disease 2019 (COVID-19), caused by severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2) is primarily a respiratory infection, which has affected the lives of millions of people all over the world. The world health organization (WHO) has announced COVID-19 a pandemic on March 2020 [1]. To detect COVID-19, several diagnostic routines based on collecting saliva or blood from the patients have been effective. However, these tests are slow and take considerable time to produce results. Cough sounds and speech based diagnosis of COVID-19 has gained interest owing to the ease of recording the signals. Recently, a database of cough sounds obtained over more than $20,000$ participants, from a range of age groups, gender, ethnicity and COVID-19 status, has been collected to facilitate detection of virus using audio signals [2]. Similar efforts are taking place in the research community [3, 4, 5]. In the speech community, two challenges as part of Interspeech 2021, namely, Interspeech 2021 ComParE challenge [6] and DiCOVA challenge [7] have been organized in that direction.

In one of the earliest studies, modeling of spectral parameters such as, spectral centroid, spectral roll–off, and zero crossing rate along with MFCCs and functionals in a Recurrent Neural Network (RNN) based– and Long Short Term Memory (LSTM) based–framework to detect the presence of COVID-19 [8]. The system was found to yield a higher classification accuracy for cough and breathing sounds compared to speech. In the Interspeech 2021 ComParE challenge, openSMILE features were found to yield better detection when compared to deep neural network based systems on the COVID-19 Speech Sub-Challenge, while on the COVID-19 Cough Sub-Challenge data augmentation together with transfer learning using pre-trained audio networks was found to yield better detection. Klump et al [9] studied the phonetic patterns in COVID-19 speech using deep acoustic model. They observed that the distinct patterns found can not be solely attributed to COVID-19. In [10], it was found that modeling of features obtained from autoregressive predictive coding neural network together with data augmentation improves cough-based COVID-19 detection. Other directions include investigation of auditory motivated features [11], combination of different spectral feature representations [12] and modeling of breathing pattern information in cough [13].

In recent years, neural network based methods have emerged which can learn information in a task dependent manner from raw speech waveform directly [14, 15, 16, 17, 18]. In this paper, we question: whether embeddings of such pre-trained neural networks without any form of adaptation can be effectively employed for COVID-19 detection? If successful, such methods can potentially serve as alternate means of finding representations that discriminate between COVID and non-COVID speech, while providing some form of explainability through the tasks on which those networks are trained. More precisely, as part of the second DiCOVA (DiCOVA-II) challenge [19], we investigate modeling of embeddings learned by neural networks trained (a) to classify phones and (b) to estimate breathing patterns, and compare them against modeling of hand-crafted paralinguistic features, namely, ComParE low level descriptors (LLDs) which have been found useful for COVID-19 detection [6, 20, 21]. We also analyze the top ranking LLDs and relate them to the information captured by the raw waveform neural networks.

The rest of the paper is organized as follows. Sec. 2 describes the features derived and classification framework utilized for our proposed method. Sec. 3 describes our experimental setup. We further present our results and analysis in Sec. 4. Sec. 5 presents the conclusion to the paper.

2. PROPOSED METHOD

Fig. 1 illustrates the proposed neural embeddings based approach. In this approach, frame level neural embeddings are extracted from pre-trained neural networks. A fixed length utterance-level representation is obtained from these embeddings, either by computing the functionals that derive first order and second order moments, or by obtaining a bag-of-audio-word (BoAW) representation. The

This work was partially funded by the Swiss National Science Foundation through the project Towards Integrated processing of Physiological and Speech signals (TIPS), grant no. 200021_188754.

fixed length representation is finally classified by using an ensemble classifier. The selection of the ensemble classification techniques was conducted in a similar manner, as presented by one of the best performing techniques during Interspeech 2020 ComParE challenge [22]. This enables us to compare in a systematic manner the neural embeddings against ComParE LLD representations.

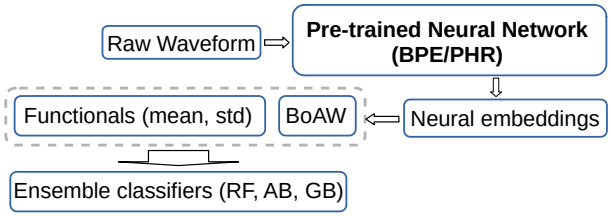

Fig. 1: The proposed neural embedding-based method for COVID-19 detection. Mean and std denote the first order and second order moments used as functionals. RF denotes Random forest [23], AB denotes Ada Boost [24] and GB denotes Gradient Boosting [25].

As mentioned in Sec. 1, we investigate neural embeddings extracted from,

1. Convolutional neural networks (CNNs) trained to model raw waveform for the task of phone classification in the context of speech recognition. One of the motivations behind using such an embedding is that COVID-19 infection affects speech production. As pointed earlier, in [9] it was found that there exist distinct phonetic patterns in COVID-19 infected speech. Although the authors conclude that those patterns may not solely attribute to COVID-19 infection, it is still worth pursuing the idea.
2. CNNs trained to model raw waveform for breathing pattern estimation. The main motivation behind that is that COVID-19 infection can adversely affect the functioning of respiratory system. As respiration process is intrinsic to speech production, breathing pattern information could be useful. In [13], such idea was pursued with modeling of estimated breathing patterns from cough in an encoder-decoder framework. In this work, we do not model the output breathing patterns but rather we model the neural embeddings extracted from an intermediate layer.

3. EXPERIMENTAL SETUP

In this section, we first present the DiCOVA-II challenge dataset and the experimental protocols. Next, we present the extraction of different fixed length representations, and finally the classifiers trained to detect COVID-19 infection.

3.1. Database and protocols

The data for the DiCOVA-II challenge [19] is derived from the Coswara dataset [3]. Speech, cough, and breathing sound recordings from 956 subjects are organized in a five fold cross–validation setting for development studies. Among these, 172 subjects were reported as tested positive for COVID-19 with mild to moderate symptoms or asymptomatic while the remaining 773 were reported as healthy with symptoms such as cold, cough, or fever, or with pre–existing respiratory conditions such as asthma. In addition, a blind test fold with 471 audio segments is provided to evaluate and report the performance of systems realized for the challenge.

The data is organized for four separate Tracks, following the same protocol. The data for the first Track includes 4.6 hours of breathing sound recordings. The second Track includes 1.7 hours of cough sound recordings, and the third Track includes 3.9 hours of speech recordings. There is no independent data for the forth Track and fusions of the systems from either of the previous Tracks is considered for evaluating the performance over this Track.

The audio and speech signals in the dataset were resampled at 16 kHz, to derive features for our experiments. We used the pre–designed protocol across five folds to report our system on the development (Dev) set. We further accumulate the data across all folds for training to evaluate our system on the Test set.

3.2. Extraction of fixed-length feature representations

We developed systems with different feature representations, namely,

1. Interspeech–2013 ComParE LLDs-based: The modeling of the LLDs was recently investigated as part of Interspeech–2021 ComParE sub-challenges for COVID-19 detection [6]. ComParE LLDs set with 65 frame level features and their 65 Δ coefficients, denoted as CMP_L, are extracted. Fixed length representations are obtained by two methods: (a) by applying functionals over low–level acoustic descriptors (CMP_L) related to energy–, spectral behaviour–, and voicing–based information, resulting in 6373 dimensional fixed length feature vector denoted as CMP_F and (b) by extracting a BoAW representation of CMP_L with two separate codebooks size of 50; the first for static and the second for Δ LLDs combined 100 dimensional representation denoted as $BoAW(CMP_L)$. The features are normalized to have zero median value. This normalization is done using the median and interquartile range between the 1st quartile (25th quantile) and the 3rd quartile (75th quantile). After this normalization, the statistical outliers are removed. We investigated these features for Track 1, Track 2 and Track 3.
2. Embedding from phone recognition neural network: we used an off-the-shelf CNN-based neural network that models raw waveform to classify phones. This network consists of ten convolutional layers followed by one hidden layer with 1024 nodes and an output layer. It was originally trained on AMI corpus [26]. Frame-level neural embeddings of 1024 dimensions, denoted as PHR, are extracted before activations of hidden layer and two different length representations are obtained (a) by computing functionals (mean and standard deviation) of the frame-level neural embeddings denoted as $f_{\mu\sigma}$(PHR), and (b) by extracting a BoAW representation of PHR with a codebook size of 100 denoted as $BoAW$(PHR). We investigated these features only on Track 2 and Track 3, as Track 1 only contains breathing recordings.
3. Embedding from breathing pattern estimation neural network: we used an off-the-shelf CNN that estimates breathing pattern at the output by taking three seconds of speech signal as input. This network consists of four convolution layers, one hidden layer with 10 nodes and one output unit. This network was originally trained on the Philips database [27] with mean squared error loss. More details about the network can be found in [18, 28]. Frame-level neural embeddings of 10 dimensions, denoted as BPE, are extracted before activations of hidden layer and two different length representations were obtained (a) by computing functionals (mean and standard deviation) of the frame-level neural embeddings denoted as $f_{\mu\sigma}$(BPE), and (b) by extracting a BoAW representation of

BPE with a codebook size of 100 denoted as $BoAW$(BPE). We investigated these features only on Track 3 for the following reasons. Track 1 contains only breathing recording. Track 2 contains cough recordings. In a previous investigation carried out as part of the Interspeech 2021 ComParE challenge, we observed that the breathing pattern during cough is considerably different from the breathing pattern during speech production. Further investigation was needed to ascertain the utility of the information extracted.

We used openSMILE toolkit [29] for extraction of CMP_L and CMP_F. We used openXBOW toolkit [30] for BoAW representation generation.

3.3. Classification

We used ensemble classification technique to train a classifier. Random Forest (RF) [23], Ada Boost (AB) [24], and Gradient Boost (GB) [25] were the ensemble classifiers used in our studies. To select the most robust ensemble classification technique, the grid search methodology, with AUC as optimization criterion, integrated in the *Scikit-learn* [31] toolkit, was used.

Tuning of hyperparameters for ensemble based classifiers was performed over the Train and Dev folds defined as per the challenge protocol. During grid search for classifier optimization, the following parameters were tuned for RF: number of estimators {500, 1000, 2000}, maximal number of features {"auto", "sqrt", "log2"}, criterion {"gini", "entropy"}, and minimal samples leaf {1, 2, 4}. For most of the cases, RF classifier yielded the best performance. The AB classifier gave comparable yet lower performance for experiments on the Dev set, and hence the results in the next Sec. 4 are presented only for RF classifier.

In addition to the framework with standalone features and classifiers, two fusion methods were also implemented to improve upon individual scores, namely, Early Fusion (EF) which is feature level combination of fixed length representations within a classification framework, and Late Fusion (LF), where the output probabilities from different systems are averaged before making decision.

4. RESULTS AND ANALYSIS

Evaluation scores for our best performing systems for different tracks of the DiCOVA-II challenge are presented in Tab. 1. As per the challenge protocol, the metrics used for evaluation are the AUC and the sensitivity on the Test set in percentage (%). The reported sensitivity is obtained at a specificity of 95%. For each track, the results for the given baseline system in the challenge is also reported. The baseline classification system uses a bidirectional LSTM (BLSTM) network to model log Mel spectrogram [19].

Our system for Track 1, based on the ComParE features, performs comparable to the baseline across the Dev set. For the Test set, the best performing system is realized by a LF of RF scores obtained using two sets of ComParE features. Even though the AUC on the Test set is lower compared to the reported baseline system, our proposed system obtains considerably better sensitivity.

On Track 2 and Track 3, the ComParE feature based systems give lower performance when compared to neural embeddings PHR based systems. PHR embeddings yield the best systems. In terms of fixed length representations, for PHR on Track 3, we observe that although both functionals and BoAW representations yield similar performance (also see Fig. 3), BoAW representation yields better sensitivity. A late fusion of the scores obtained with the two representations marginally increases the AUC on the Test set, however does not contribute to the sensitivity. Similar observations can be noted for Track 2 except that late fusion of these representations slightly increases both AUC and sensitivity on the test set.

Table 1: Results obtained for different systems over Dev and Test set of the DiCOVA-II challenge. The results are expressed in AUC metric and the sensitivity of the systems on the Test set at specificity of 95%. The systems noted as *[I]*, *[II]*, *[III]*, and *[IV]* were used in fusion method for Track 4.

System Feature	Classifier	Dev (%)	Test (%)	Sensitivity (%)
Track 1				
CMP_F	RF	77.83	76.78	30.0
$BoAW(CMP_L)$	RF	73.58	74.52	31.67
CMP_F , $BoAW(CMP_L)$ ***[I]***	LF	77.56	**78.05**	**43.33**
BASELINE	BLSTM	77.25	84.50	31.67
Track 2				
$BoAW$(PHR)	RF	70.06	74.19	30.0
$f_{\mu\sigma}$(PHR)	RF	70.54	72.87	26.67
CMP_L	RF	66.09	66.68	16.67
$f_{\mu\sigma}$(PHR), $BoAW$(PHR) ***[II]***	LF	71.32	**74.63**	**31.67**
BASELINE	BLSTM	75.21	74.89	36.67
Track 3				
$BoAW$(PHR) ***[III]***	RF	77.37	80.08	**41.67**
$f_{\mu\sigma}$(PHR)	RF	76.33	79.3	26.67
$BoAW$(BPE)	RF	68.93	73.49	21.67
$f_{\mu\sigma}$(BPE)	RF	68.44	—	—
$BoAW(CMP_L)$	RF	70.38	75.59	15.0
EF($f_{\mu\sigma}$(PHR), $f_{\mu\sigma}$(BPE)) ***[IV]***	RF	76.67	79.1	28.33
EF($BoAW$(PHR), $BoAW$(BPE) , $BoAW(CMP_L)$)	RF	77.47	79.95	33.33
$f_{\mu\sigma}$(PHR), $BoAW$(PHR)	LF	77.59	**80.64**	36.67
BASELINE	BLSTM	80.16	84.26	43.33
Track 4				
[III], [IV]	LF	77.79	**80.51**	40.0
[I], [IV]	LF	80.09	78.05	**43.33**
[I], [II]	LF	77.93	78.05	**43.33**
BASELINE	LF	81.67	84.70	55.0

Looking into the system performances for Track 4, it appears that the classifiers trained using breathing sounds are more prominent when fused with systems trained with cough and speech signals. The fused systems in Track 4 has higher AUC compared to system from Track 2 and comparable AUC to the best system from Track 3.

4.1. Analysis of neural embedding based systems

Figure 2 shows the cumulative frequency response of the first convolutional layer for PHR CNN and BPE CNN. It can be observed that the PHR network emphasizes around the formant frequency regions in speech, while the emphasis of the BPE network is significantly towards the lower frequency region. In other words, they are modeling different information from the speech signal.

Fig. 3 shows the ROC plot for both feature sets of PHR– and BPE– based networks for Track 3. It can be seen that the PHR neural embeddings consistently yield better system than BPE neural embeddings. One of the reason for that could be that the participants may not have had severe COVID issues for the differences w.r.t non-COVID participants to be very apparent at BPE embedding level.

4.2. Analysis of LLD-based systems

In order to analyze the discriminability of the features used for our studies, we estimated their respective importance in achieving the

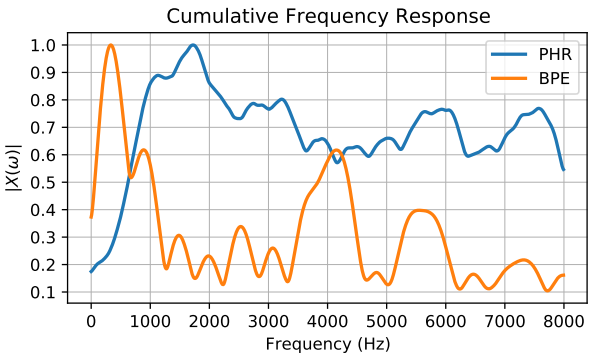

Fig. 2: The cumulative frequency response of the kernels for the first convolution layer of the CNN models: PHR and BPE.

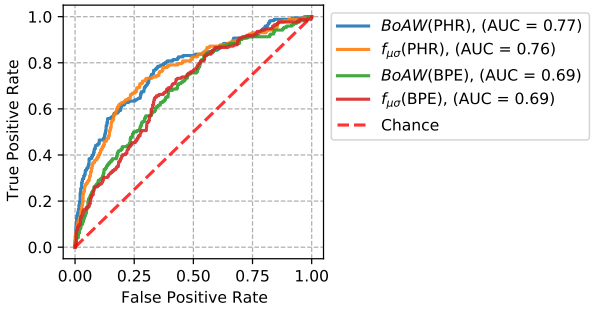

Fig. 3: ROC plot for systems trained using PHR embeddings and BPE embeddings on the Dev set of Track 3.

desired classification performance. This analysis improves our comprehension towards the significance of a certain feature for identifying COVID-19 positive cases in a given audio modality. Tab. 2 presents an overview of the most important features for each track, based on an importance scores generated by the RF-based classifier.

For all the tracks, the auditory spectra coefficients obtained using RASTA filtering (audSpec:Rfilt) and their deltas (Δ) establish as one of the most discriminative LLDs. For Track 1, coefficients obtained as the third quartile of these features prove significant for classification. For Track 2, an extended list of functionals prove significant with features capturing primarily the spectral shape. For Track 3, speech specific features such as MFCC and spectral band energy prove discriminatory for the task.

5. SUMMARY AND FUTURE WORK

In this paper, we investigated modeling of neural embeddings extracted from raw waveform modeling neural networks, pre-trained for phone classification and breathing pattern estimation, for the task of COVID-19 infection detection. More precisely, these embeddings were modeled as fixed length representations through application of functionals and BoAW, similar to modeling of hand-crafted LLDs for paralinguistic speech processing. Our investigations on DiCOVA-II challenge showed that neural embeddings extracted from phone classification neural network, i.e. PHR, can yield better systems than hand-crafted LLD-based systems and BPE embedding-based systems. On Track 3, we observed that, although BPE embedding-based system yields slightly lower performance than LLD-based system, it yields considerably better sensitivity. Taken together our studies demonstrate that modeling neural embeddings from neural networks trained on auxiliary or other speech tasks for COVID-19 infection detection is a promising direction and can replace hand-crafted features.

In our studies, despite the neural networks being trained on auxiliary data and tasks, we found that the proposed neural embedding based systems were comparable to the baseline system on Track 2 and somewhat inferior to the baseline system in Track 3. This suggests that there is room for further improvements.

Our future work will focus in that direction: (a) by investigating adaptation of the investigated pre-trained neural networks on target data for COVID-19 detection and (b) by investigating pre-trained neural networks that focus on other aspects of speech production such as, voice source [32].

Table 2: LLDs and functionals exhibiting highest discriminability for each track (most representative, non-redundant features).

LLDs	functional
Track 1	
Δ audSpec_Rfilt	3^{rd} quartile
voicing parameters	LP–gain
magnitude spectra	RollOff
Δ magnitude spectra	variance
Track 2	
audSpec_Rfilt	regression coefficients, centroid, 2^{nd} quartile
Δ Pitch contour	regression coefficients
Δ RMSenergy	extremums
band energy magnitude spectra	extremums
magnitude spectral slope	regression coefficients
Track 3	
audSpec_Rfilt	regression coefficients, 1^{st} quartile
mfcc	peak behavior , percentiles
Δ audSpec_Rfilt	peak behavior
Δ magnitude spectra	moments

6. REFERENCES

[1] "Who," https://www.who.int/health-topics/coronavirus.

[2] Lara Orlandic et al., "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.

[3] Neeraj Sharma et al., "Coswara — A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis," in *Proceedings of Interspeech*, 2020, pp. 4811–4815.

[4] Chloë Brown et al., "Exploring automatic diagnosis of covid-19 from crowdsourced respiratory sound data," in *International Conference on Knowledge Discovery & Data Mining (ACM SIGKDD)*, New York, NY, USA, 2020, p. 3474–3484, Association for Computing Machinery.

[5] Jing Han et al., "Exploring automatic covid-19 diagnosis via voice and symptoms from crowdsourced data," in *ICASSP*, 2021, pp. 8328–8332.

[6] B. W. Schuller et al., "The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-

19 Speech, Escalation & Primates," in *Proceedings of Interspeech*, 2021, pp. 431–435.

[7] Ananya Muguli et al., "DiCOVA Challenge: Dataset, Task, and Baseline System for COVID-19 Diagnosis Using Acoustics," in *Proceedings of Interspeech*, 2021, pp. 901–905.

[8] Abdelfatah Hassan et al., "Covid-19 detection system using recurrent neural networks," in *International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI)*. IEEE, 2020, pp. 1–5.

[9] P. Klumpp et al., "The Phonetic Footprint of Covid-19?," in *Proceedings of Interspeech*, 2021, pp. 441–445.

[10] John Harvill et al., "Classification of COVID-19 from Cough Using Autoregressive Predictive Coding Pretraining and Spectral Data Augmentation," in *Proceedings of Interspeech*, 2021, pp. 926–930.

[11] Rohan Kumar Das et al., "Diagnosis of COVID-19 Using Auditory Acoustic Cues," in *Proceedings of Interspeech*, 2021, pp. 921–925.

[12] Kotra Venkata Sai Ritwik et al., "COVID-19 Detection from Spectral Features on the DiCOVA Dataset," in *Proceedings of Interspeech*, 2021, pp. 936–940.

[13] Gauri Deshpande et al., "The DiCOVA 2021 Challenge — An Encoder-Decoder Approach for COVID-19 Recognition from Coughing Audio," in *Proceedings of Interspeech*, 2021, pp. 931–935.

[14] Dimitri Palaz et al., "Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks," in *Proceedings of Interspeech*, 2013, pp. 1766–1770.

[15] George Trigeorgis et al., "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proceedings of ICASSP*, 2016, pp. 5200–5204.

[16] Hannah Muckenhirn et al., "Towards directly modeling raw speech signal for speaker verification using cnns," in *Proceedings of ICASSP*, 2018, pp. 4884–4888.

[17] H. Muckenhirn et al., "Understanding and visualizing raw waveform-based CNNs," in *Proceedings of Interspeech*, 2019, pp. 2345–2349.

[18] V. S. Nallanthighal et al., "Deep learning architectures for estimating breathing signal and respiratory parameters from speech recordings," *Neural Networks*, vol. 141, pp. 211–224, 2021.

[19] "Second dicova challenge," `https:https://dicovachallenge.github.io/`.

[20] Flavio Avila et al., "Investigating Feature Selection and Explainability for COVID-19 Diagnostics from Cough Sounds," in *Proceedings of Interspeech*, 2021, pp. 951–955.

[21] I. Södergren et al., "Detecting COVID-19 from Audio Recording of Coughs Using Random Forests and Support Vector Machines," in *Proceedings of Interspeech*, 2021, pp. 916–920.

[22] Maxim Markitantov et al., "Ensembling End-to-End Deep Models for Computational Paralinguistics Tasks: ComParE 2020 Mask and Breathing Sub-Challenges," in *Proceedings of Interspeech*, 2020, pp. 2072–2076.

[23] Tin Kam Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*. IEEE, 1995, vol. 1, pp. 278–282.

[24] Yoav Freund et al., "Experiments with a new boosting algorithm," in *icml*. Citeseer, 1996, vol. 96, pp. 148–156.

[25] Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Frean, "Boosting algorithms as gradient descent in function space," in *Proc. NIPS*, 1999, vol. 12, pp. 512–518.

[26] Jean Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus," *Language Resources and Evaluation*, vol. 41, no. 2, pp. 181–190, 2007.

[27] V. S. Nallanthighal et al., "Deep sensing of breathing signal during conversational speech," in *Proceedings of Interspeech*. Graz, Austria, 2019, pp. 4110–4114.

[28] Zohreh Mostaani et al., "On the relationship between speech-based breathing signal prediction evaluation measures and breathing parameters estimation," in *Proceedings of ICASSP*, 2021, pp. 1345–1349.

[29] Florian Eyben et al., "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.

[30] Maximilian Schmitt et al., "openxbow–introducing the passau open-source crossmodal bag-of-words toolkit," *Journal of Machine Learning Research*, vol. 18, no. 96, pp. 1–5, 2017.

[31] Fabian Pedregosa et al., "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[32] S. P. Dubagunta et al., "Learning voice source related information for depression detection," in *Proceedings of ICASSP*, 2019, pp. 6525–6529.