COVID-19 Detection Exploiting Self-Supervised Learning Representations of Respiratory Sounds

Adria Mallol-Ragolta*†, Shuo Liu*, and Björn Schuller*†‡

* *EIHW – Chair of Embedded Intelligence for Health Care & Wellbeing, University of Augsburg, Germany*
† *Centre for Interdisciplinary Health Research, University of Augsburg, Germany*
‡ *GLAM – Group on Language, Audio, & Music, Imperial College London, UK*

adria.mallol-ragolta@uni-a.de

***Abstract*—In this work, we focus on the automatic detection of COVID-19 patients from the analysis of cough, breath, and speech samples. Our goal is to investigate the suitability of Self-Supervised Learning (SSL) representations extracted using Wav2Vec 2.0 for the task at hand. For this, in addition to the SSL representations, the models trained exploit the Low-Level Descriptors (LLD) of the eGeMAPS feature set, and Mel-spectrogram coefficients. The extracted representations are analysed using Convolutional Neural Networks (CNN) reinforced with contextual attention. Our experiments are performed using the data released as part of the Second Diagnosing COVID-19 using Acoustics (DiCOVA) Challenge, and we use the Area Under the Curve (AUC) as the evaluation metric. When using the CNNs without contextual attention, the multi-type model exploiting the SSL Wav2Vec 2.0 representations from the cough, breath, and speech sounds scores the highest AUC, 80.37 %. When reinforcing the embedded representations learnt with contextual attention, the AUC obtained using this same model slightly decreases to 80.01 %. The best performance on the test set is obtained with a multi-type model fusing the embedded representations extracted from the LLDs of the cough, breath, and speech samples and reinforced using contextual attention, scoring an AUC of 81.27 %.**

***Index Terms*—COVID-19 Detection, Respiratory Diagnosis, Paralinguistics, Self-Supervised Representations, Healthcare**

I. Introduction

Digital health technologies based on *Artificial Intelligence* (AI) can be used to develop large-scale, cost-effective solutions for massive population screenings that ultimately contribute to the early detection of diseases. In the current *Coronavirus Disease 2019* (COVID-19) pandemic context, such solutions could be used as a pre-screening tool to reduce the number of medical tests performed, which are expensive, time-consuming, and generate a large amount of waste. Previous works have explored the use of AI-based solutions in a wide range of health-related problems, including the recognition of mental illnesses, such as depression [1], [2] or *Post-Traumatic Stress Disorder* (PTSD) [3]. Motivated by the pandemic, recent works have focused on the detection of COVID-19 from the analysis of different modalities, including respiratory sounds [4]–[6].

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826506 (sustAGE), and from the DFG's Reinhart Koselleck project No. 442218748 (AUDI0NOMOUS). It has also been supported by the Bavarian Ministry of Science and Arts through the ForDigitHealth project, funded by the Bavarian Research Association on Healthy Use of Digital Technologies and Media.

The COVID-19 research has not yet determined the most suitable features to extract from the respiratory sounds. Inspired by the promising performances of *Self-Supervised Learning* (SSL) representations [7], [8], we aim to investigate –to the best of the authors' knowledge, for the first time– the extraction of SSL representations from coughs, breaths, and speech for the automatic detection of COVID-19 patients. To enrich the comparison of feature representations, we include the *Low-Level Descriptors* (LLD) of the eGeMAPS feature set [9], and Mel-spectrogram coefficients in our analysis.

We use the dataset released as part of the Second *Diagnosing COVID-19 using Acoustics* (DiCOVA) Challenge [10], [11] to assess the performance of mono-type and multi-type models exploiting the aforementioned feature representations. The networks implemented to extract the salient information from the extracted feature representations are composed of two main blocks: the first block extracts embedded representations from the input features, while the second block is responsible for the actual classification. Furthermore, we explore the use of contextual attention [1], [12] with the goal of learning to highlight the embedded representations obtained at the output of the first block that contribute the most to the task.

The rest of this paper is laid out as follows: Section II describes the dataset analysed, while Section III details the methodology followed. Section IV compiles and analyses the results obtained, and Section V concludes the paper.

II. Dataset

The dataset released as part of the Second DiCOVA Challenge [10], [11] contains cough, breath, and speech samples –which are sound types produced by the human respiratory system– from COVID-19 positive and negative (healthy) patients. The sampling rate of the respiratory sounds provided by the Challenge organisers is 44.1 kHz. However, a preliminary exploration of the dataset revealed that some samples do not contain frequency information in the upper frequencies of the spectrogram. This observation suggests that some audio samples were originally recorded at a different, lower sampling rate, and upsampled before distributing the data. This is a plausible hypothesis given the nature of the dataset, which was recorded in-the-wild, via crowdsourcing, and using the patients' own devices. The available samples are distributed in

TABLE I
DATA AVAILABLE IN THE SECOND DICOVA CHALLENGE DATASET TIME-WISE PER SOUND TYPE AND DATA PARTITION. THE TEMPORAL INFORMATION IS PROVIDED IN THE FORMAT (HH:)MM:SS.

(HH:)MM:SS	**Validation**	**Test**	$\sum$
Cough	1:41:01	37:58	2:18:59
Breath	4:37:37	2:07:46	6:45:23
Speech	3:56:22	1:44:39	5:41:01
$\sum$	10:15:00	4:30:23	14:45:23

two partitions, and the Challenge organisers require assessing the performance of the models on the training partition using a pre-defined 5-fold cross-validation approach.

Each patient recorded a cough, a breath, and a speech sample. The total duration of the dataset is 14 h 45 min 23 sec (cf. Table I). The dataset contains information from a total of 1 436 patients (cf. Table II): 965 belonging to the training partition, and 471, to the test partition. The training data is imbalanced both in terms of sex (242 females and 723 males) and COVID-19 status (172 positives and 793 negatives). Similarly, the test data is also imbalanced in terms of sex (119 females and 352 males), whilst the COVID-19 status distribution is blind to the Challenge participants.

III. METHODOLOGY

This section presents the methodology. Section III-A details the pre-processing applied to the audio samples, Section III-B describes the networks implemented, Section III-C indicates the post-processing applied to the model inferences, and Section III-D summarises the network training parameters.

A. Data Preparation

This section introduces the data conditioning applied to the respiratory sounds. Section III-A1 details the pre-processing applied to the cough, breath, and speech samples, while Section III-A2 describes the different feature representations extracted from the respiratory sounds.

1) Respiratory Sounds Pre-Processing: The dataset explored was collected in-the-wild, using the patients' own recording devices (cf. Section II). To overcome this disparity, the respiratory sounds are first converted to 16 kHz and mono-channel. After listening to a subset of the recordings, we detected that: i) the coughs, the breaths, or the speech might start and finish a few seconds after and before the actual start and end of the recordings, respectively, and ii) some recordings are empty and do not contain respiratory sounds.

The former can be attributed to the self-recording procedure implemented to collect the data, as patients might have needed a preparation phase of a few seconds to start and stop the recording before starting and finishing coughing, breathing, or speaking. To exclude the information unintentionally collected during the preparation phase, we implement a *Root-Mean-Square* (RMS) energy-based *Sound Activity Detector* (SAD). The RMS features are computed using a frame length of 1 024 samples (64 ms) and a hop length of 512 samples (32 ms). The resulting RMS signal is normalised, and we

TABLE II
STATISTICS OF THE SECOND DICOVA CHALLENGE DATASET IN TERMS OF THE PATIENTS' SEX AND THEIR COVID-19 STATUS. THE LATTER IS BLIND TO THE CHALLENGE PARTICIPANTS ON THE TEST SET.

#	**Validation**			**Test**	$\sum$
	Positive	Negative	$\sum$		
Females	53	189	242	119	361
Males	119	604	723	352	1 075
$\sum$	172	793	965	471	1 436

experimentally define a threshold of 0.1 to differentiate the content-rich frames from the silent ones. The timestamps of the first and last frames whose RMS-based energy is above the threshold are used to segment the original respiratory sound.

The empty recordings have a short duration –usually, below 2 sec– and mainly contain background noise. To automatically identify these samples and exclude them from the training process, we check the mean RMS-based energy of the acoustic signals whose duration is shorter than 2 sec. Considering the nature of the energy signal computed from a prototypical speech sample, which contains voiced and unvoiced frames, we empirically define a threshold of 0.5 to differentiate when a respiratory sound contains relevant information from background noise. If the mean RMS-based energy is above the threshold, we consider the corresponding sample as empty. The cough, breath, and speech samples of a patient are considered valid for training if and only if none of them is interpreted as empty by the described procedure. Following this approach, 3 patients from the validation partition are excluded from the training material for providing empty respiratory sounds.

Next, we homogenise the duration of the cough, breath, and speech samples from each individual patient to ease the fusion of the different sound types. For this, we determine the longest respiratory sound recorded by each patient and extend via repetition the shorter ones. Despite this intra-patient homogenisation, the duration of the respiratory sounds is patient-dependent. As sequences of the same length are commonly used to train neural networks, we decide to model the acoustic information in windows of 5 sec. In case the length of the respiratory sounds recorded by a patient, even after the aforementioned homogenisation, are shorter than 5 sec, these are all extended via repetition, so at least one window of information can be computed from each respiratory sound.

2) Features Extraction: In this work, we aim to compare the model performances when exploiting the 25-dimensional LLDs of the eGeMAPS feature set [9] extracted using OPENSMILE [13], 128-dimensional Mel-spectrogram coefficients, and the 768-dimensional SSL features extracted with the pre-trained Wav2Vec 2.0 base model [14]. Each feature representation has a different resolution in the time domain: the LLDs are extracted at 100 Hz, the Mel-spectrogram coefficients at 125 Hz, and the Wav2Vec 2.0 representations at 50 Hz. Finally, we window the extracted representations separately without overlap, so each segment contains the features corresponding to 5 sec of the pre-processed respiratory sounds.

TABLE III

AUC MEASUREMENTS (%) OBTAINED FROM THE MONO- AND MULTI-TYPE MODELS TRAINED EXPLOITING THE LLDS OF THE EGEMAPS FEATURE SET, THE MEL-SPECTROGRAM COEFFICIENTS, AND THE SSL WAV2VEC 2.0 REPRESENTATIONS EXTRACTED FROM C(OUGHS), B(REATHS), AND S(PEECH).

Sound types	Set	LLDs of eGeMAPS	Mel-Spec. Coeff.	Wav2Vec 2.0
C	Val.	70.43	72.12	56.21
	Test	71.30	69.15	58.22
B	Val.	72.16	76.18	67.79
	Test	62.22	74.26	65.02
S	Val.	72.13	72.04	71.86
	Test	74.01	57.43	76.73
$C \oplus B$	Val.	72.17	74.14	67.16
	Test	**79.07**	74.25	65.43
$C \oplus S$	Val.	71.67	72.05	66.92
	Test	69.08	71.87	74.88
$B \oplus S$	Val.	74.56	74.06	72.27
	Test	77.19	**76.48**	73.18
$C \oplus B \oplus S$	Val.	72.84	74.77	69.44
	Test	79.05	76.30	**80.37**

(a) Models using CNNs to learn the embedded representations from the features.

Sound types	Set	LLDs of eGeMAPS	Mel-Spec. Coeff.	Wav2Vec 2.0
C	Val.	71.03	71.44	57.89
	Test	66.03	65.07	58.05
B	Val.	72.22	75.42	69.48
	Test	78.89	78.43	69.95
S	Val.	71.60	72.03	71.15
	Test	77.29	67.99	76.03
$C \oplus B$	Val.	73.12	74.00	67.76
	Test	78.71	74.17	69.59
$C \oplus S$	Val.	71.59	73.57	68.65
	Test	77.48	71.75	75.88
$B \oplus S$	Val.	73.28	74.99	74.31
	Test	80.40	**80.46**	79.98
$C \oplus B \oplus S$	Val.	73.58	74.54	70.60
	Test	**81.27**	79.28	**80.01**

(b) Models using CNNs reinforced with contextual attention to learn the embedded representations from the features.

B. Models Description

The networks implemented in this work are composed of 2 main blocks: the first block extracts embedded representations from the extracted feature representations (cf. Section III-A), while the second block performs the actual classification. The classification block implements two *Fully Connected* (FC) layers with 32 and 2 output neurons, respectively, both preceded by dropout layers with probability 0.3. While the outputs of the first FC layer are transformed using a *Rectified Linear Unit* (ReLU) activation function, the second FC layers uses Softmax as the activation function, so the network outputs can be interpreted as probability scores.

The embeddings extraction block implements specific *Convolutional Neural Networks* (CNN) to extract embedded representations from the cough, breath, and speech representations separately. The mono-type models have a single embeddings extraction block, while the multi-type models fusing the cough, breath, and speech samples contain 3 specific embeddings extraction blocks. Each embeddings block implements 2 1-dimensional CNN layers with 64, and 128 output filters, respectively, using a kernel size of 3, and a stride of 1. Following each convolutional layer, we use batch normalisation and transform the outputs using a ReLU function. A 1-dimensional max-pooling layer, and a 1-dimensional adaptive average pooling layer are included at the end of the first, and second convolutional blocks, respectively. The multi-type models fuse the embedded representations learnt at the output of the embeddings extraction blocks via concatenation. The dimensionality of the input features to the classification block depends on the number of sound types to be fused together.

Additionally, we aim to reinforce the embedded representations learnt with contextual attention [1], [12], so we can analyse how this mechanism impacts the performance of the different feature representations. The contextual attention mechanisms should help highlight the salient information in the embedded representations learnt, and, therefore, we implement a specific contextual attention mechanism to each embeddings extraction block. The attention-based representation obtained is then fed into the classification block of the network when training mono-type models, or fused via concatenation when training multi-type models.

C. Inferences Post-Processing

Because of the windowing procedure described in Section III-A, several instances of feature representations may be extracted from the same patient, and, consequently, several probability scores may be inferred (one from each window). In these cases, we collect all the model inferences corresponding to the same patient and define the mean of the individual probabilities as the final probability inferred by the model.

D. Networks Training

The available data is imbalanced (cf. Section II), which negatively impacts the learning capabilities of the models. To overcome this issue, we implement a weighted random sampler to select the training samples to use in each batch of the training routine, so they are balanced in terms of their COVID-19 status. All models are trained under the exact same conditions for a fair comparison. We define the Categorical Cross-Entropy as the loss to minimise, using Adam as the optimiser with a fixed learning rate of 10^{-3}. As model performances are assessed in terms of the *Area Under the Curve* (AUC), we define $\mathcal{L}_{\mathrm{AUC}} = 1 - \mathrm{AUC}$ as the validation loss to monitor during the training process. Network parameters are updated in batches of 64 samples, and trained during a maximum of 100 epochs. We implement an early-stopping mechanism to stop training when the validation loss does not improve for 15 consecutive epochs. We follow a 5-fold cross-validation approach to evaluate the models. As each fold is trained during a specific number of epochs, when modelling all training material and to prevent overfitting, the training epochs are determined by computing the mean of the training epochs processed in each fold, rounded up to the next integer.

IV. Experimental Results

The results from the mono- and the multi-type models without and with contextual attention are summarised in Table III.

Comparing the performance of the mono-type models (cf. Table III.a), we observe that the highest AUC scores on the test set when using cough, breath, and speech samples are obtained when exploiting the LLDs of eGeMAPS, the Mel-spectrogram coefficients, and the SSL Wav2Vec 2.0 representations, respectively, with an AUC of 71.30 %, 74.26 %, and 76.73 %. The underperformance of the SSL Wav2Vec 2.0 representations on the cough and breath samples can be attributed to the nature of the representations, as Wav2Vec 2.0 is pre-trained on a speech dataset. When the networks do not use contextual attention at the output of the embeddings extraction block, the highest AUC on the test set is achieved by the multi-type model fusing the SSL Wav2Vec 2.0 representations of the cough, breath, and speech samples, with an AUC of 80.37 %.

Comparing the performance of the mono-type models when reinforcing the outputs of the embeddings extraction block with contextual attention (cf. Table III.b), we observe that the highest AUC scores on the test set when using cough, breath, and speech samples are obtained with the LLDs of eGeMAPS, with an AUC of 66.03 %, 78.89 %, and 77.29 %. The results obtained with the multi-type models suggest that when cough samples are involved in the fusion, the most suitable feature representations to extract are the LLDs of eGeMAPS. To fuse the breath and speech samples, the most suitable features to exploit are the Mel-spectrogram coefficients. When the networks do use contextual attention at the output of the embeddings extraction block, the highest AUC on the test set is achieved by the multi-type model fusing the LLDs of eGeMAPS extracted from the cough, breath, and speech samples, with an AUC of 81.27 %. In this scenario, the best multi-type model exploiting SSL Wav2Vec 2.0 representations scores the third position, behind the LLDs of eGeMAPS, and the Mel-spectrogram coefficients. This result seems to cast doubt on the use of contextual attention on top of a transformer-based network –as Wav2Vec 2.0 essentially is–, which implementation is based on self-attention mechanisms.

V. Conclusions

This work compared the performance of mono- and multi-type COVID-19 detection models exploiting the LLDs of the eGeMAPS feature set, the Mel-spectrogram coefficients, and the SSL Wav2Vec 2.0 representations extracted from respiratory sounds. When contextual attention mechanisms were not used, the fusion of the SSL Wav2Vec 2.0 representations extracted from the cough, breath, and speech samples scored the highest AUC, 80.37 %. Nevertheless, the best performance on the test set was obtained with the model using contextual attention, and fusing the LLDs of the eGeMAPs feature set extracted from the coughs, breaths, and speech, 81.27 %.

To train the multi-type models, the same representations were extracted from the fused sound types. However, the results obtained from the mono-type models without contextual attention indicated that the most suitable feature representation depended on the sound type to be exploited. A follow-up study could investigate the use of different representations to characterise the sound types to be fused. Further research includes the exploration of few-shot learning to account for the scarcity of COVID-19 positive patients in the dataset.

References

[1] Adria Mallol-Ragolta, Ziping Zhao, Lukas Stappen, Nicholas Cummins, and Björn Schuller, "A Hierarchical Attention Network-Based Approach for Depression Detection from Transcribed Clinical Interviews," in *Proc. of Interspeech*, Graz, Austria, 2019, pp. 221–225, ISCA.

[2] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, Siyang Song, Shuo Liu, Ziping Zhao, Adria Mallol-Ragolta, Zhao Ren, Mohammad Soleymani, and Maja Pantic, "AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition," in *Proc. of the 9th Intl. Audio/Visual Emotion Challenge and Workshop*, Nice, France, 2019, pp. 3–12, ACM.

[3] Adria Mallol-Ragolta, Svati Dhamija, and Terrance E. Boult, "A Multimodal Approach for Predicting Changes in PTSD Symptom Severity," in *Proc. of the 20th Intl. Conf. on Multimodal Interaction*, Boulder, CO, USA, 2018, pp. 324–333, ACM.

[4] Adria Mallol-Ragolta, Helena Cuesta, Emilia Gómez, and Björn Schuller, "Cough-based COVID-19 Detection with Contextual Attention Convolutional Neural Networks and Gender Information," in *Proc. of Interspeech*, Brno, Czechia, 2021, pp. 941–945, ISCA.

[5] Adria Mallol-Ragolta, Florian B. Pokorny, Katrin D. Bartl-Pokorny, Anastasia Semertzidou, and Björn Schuller, "Triplet Loss-Based Models for COVID-19 Detection from Vocal Sounds," in *Proc. of the 44th Annual Intl. Conf. of the Engineering in Medicine & Biology Society*, Glasgow, UK, 2022, IEEE, 4 pages, to appear.

[6] Adria Mallol-Ragolta, Helena Cuesta, Emilia Gómez, and Björn Schuller, "Multi-Type Outer Product-Based Fusion of Respiratory Sounds for Detecting COVID-19," in *Proc. of Interspeech*, Incheon, Korea, 2022, ISCA, 5 pages, to appear.

[7] Amrita Bhattacharjee, Mansooreh Karami, and Huan Liu, "Text Transformations in Contrastive Self-Supervised Learning: A Review," 2022, arXiv:2203.12000.

[8] Shuo Liu, Adria Mallol-Ragolta, Emilia Parada-Cabeleiro, Kun Qian, Xin Jing, Alexander Kathan, Bin Hu, and Björn Schuller, "Audio Self-supervised Learning: A Survey," 2022, arXiv:2203.01205.

[9] Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, April 2016.

[10] Neeraj Sharma, Prashant Krishnan, Rohit Kumar, Shreyas Ramoji, Srikanth Raj Chetupalli, Nirmala R., Prasanta Kumar Ghosh, and Sriram Ganapathy, "Coswara – A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis," in *Proc. of Interspeech*, Shanghai, China, 2020, pp. 4811–4815, ISCA.

[11] 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," in *Proc. of the Intl. Conf. on Acoustics Speech Signal Processing*, Singapore, 2022, IEEE, to appear.

[12] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy, "Hierarchical Attention Networks for Document Classification," in *Proc. of the Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, CA, USA, 2016, pp. 1480–1489, ACL.

[13] Florian Eyben, Martin Wöllmer, and Björn Schuller, "openSMILE – The Munich Versatile and Fast Open-source Audio Feature Extractor," in *Proc. of the 18th Intl. Conf. on Multimedia*, Firenze, Italy, 2010, pp. 1459–1462, ACM.

[14] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Proc. of the 34th Conf. on Neural Information Processing Systems*, Vancouver, Canada, 2020, pp. 12449–12460, Curran Associates, Inc.